# Seeing Reason: Visuospatial Ability, Sex Differences and the Raven's Progressive Matrices

Nicolette Amanda Waschl

School of Psychology, University of Adelaide

*A thesis submitted in fulfillment of the requirements for the degree of*

*Doctor of Philosophy*

March, 2017

# Table of Contents

# List of Tables

# List of Figures

# List of Abbreviations

| | |
|---|---|
| 2PL | Two-parameter logistic |
| 3PL | Three-parameter logistic |
| AH | Alice Heim Test |
| APM | Raven’s Advanced Progressive Matrices |
| AR | Abstract Reasoning (GRT2 subtest) |
| BIS | Berlin Structure of Intelligence |
| CAB-I | Comprehensive Ability Battery – Inductive Reasoning |
| CAB-Cf | Comprehensive Ability Battery – Flexibility of Closure |
| CFA | Confirmatory Factor Analysis |
| CFI | Comparative Fit Index |
| CFIT | Cattell’s Culture Fair Intelligence Test |
| CHC | Cattell-Horn-Carroll |
| CPM | Raven’s Coloured Progressive Matrices |
| CTT | Classical Test Theory |
| DAS | Differential Aptitude Scales |
| DAT | Differential Aptitude Tests |
| DAT-AR | Differential Aptitude Tests – Abstract Reasoning |
| DAT-VR | Differential Aptitude Tests – Verbal Reasoning |
| DIF | Differential item functioning |
| DWLS | Diagonally weighted least squares |
| EA | Esoteric Analogies Test |
| EFA | Exploratory factor analysis |
| ESEM | Exploratory structural equation modeling |
| ETS | Educational Testing Service |

| | |
|---|---|
| *g* | General intelligence |
| Gc | Crystallised ability |
| Gf | Fluid ability |
| GRT2 | General Reasoning Test 2 |
| Gq | Quantitative ability |
| Gv | Visuospatial ability |
| I | Induction |
| ICC | Item characteristic curve |
| IQ | Intelligence Quotient |
| IRT | Item Response Theory |
| IST | Intelligence Structure Test |
| K-BIT | Kaufman Brief Intelligence Test |
| K-SNAP | Kaufman Short Neuropsychological Assessment Procedure |
| KAIT | Kaufman Adolescent and Adult Intelligence Test |
| MGCFA | Multiple-group confirmatory factor analysis |
| MIMIC | Multiple-indicator Multiple-causes |
| ML | Maximum likelihood |
| MRT | Mental Rotation Test |
| MTMM | Multitrait-multimethod |
| NR | Numerical Reasoning (GRT2 subtest) |
| PCA | Principal components analysis |
| PF | Paper Folding Test |
| PMA-R | Primary Mental Abilities – Reasoning |
| PSVT:R | Perdue Spatial Visualization Test of Rotations |
| RG | Sequential reasoning |

| | |
|---|---|
| RMSEA | Root Mean Square Error of Approximation |
| RPM | Raven’s Progressive Matrices |
| RQ | Quantitative reasoning |
| SEM | Structural Equation Modeling |
| SRMR | Standardized Root Mean Square Residual |
| SPM | Standard Raven’s Progressive Matrices |
| US | United States |
| UK | United Kingdom |
| VPR | Verbal-Perceptual-Image Rotation |
| VR | Verbal Reasoning (GRT2 subtest) |
| WAIS | Wechsler Adult Intelligence Scale |
| WLSMV | Weighted least squares mean and variance adjusted |
| WJ-III | Woodcock-Johnson Test of Cognitive Abilities – 3rd Edition |

# Abstract

This thesis sought to address the role of visuospatial ability in measures of inductive reasoning, with a particular focus on the Raven’s Progressive Matrices (RPM). Given that males tend to perform better on certain measures of visuospatial ability, sex differences in performance on the RPM tests and in other measures of inductive reasoning were also examined.

The issue of the involvement of visuospatial ability in the RPM tests is important at both a practical and a theoretical level. At the practical level, these tests are often used as a sole measure of general intelligence, and conclusions regarding the relationship of general intelligence to other variables are made on the basis of results from this test. If the RPM tests require a substantive amount of visuospatial ability, this is problematic to the interpretation of results on this test as reflective of general intelligence. At a theoretical level, investigation of this question pertains to an understanding of the relationship between visuospatial abilities and fluid ability generally, but inductive reasoning more specifically. Many commonly used measures of inductive reasoning are presented in a visual format (e.g. abstract figures) and these tests are often shown to cross-load on both fluid and visuospatial factors.

This thesis addresses the issues of visuospatial ability and sex differences in the RPM by examining (1) the dimensionality of the Advanced RPM tests; (2) the role of Gv in performance on the RPM tests; and (3) sex differences in raw scores on the RPM and other measures of inductive reasoning. Additionally, the psychometric properties of the General Reasoning Test 2 (GRT2) in the Australian population were examined. This included an investigation of the relationship between figural, verbal and numeric reasoning items as well as sex differences.

Study 1 used confirmatory factor analysis and Rasch modeling to investigate the dimensionality of the Advanced RPM, measurement invariance and differential item functioning across sex. Study 2 used structural equation modeling to examine, in three separate samples, how well visuospatial abilities could account for the variance in a latent RPM factor not already accounted for by alternative fluid ability measures. This study additionally assessed invariance of the structural relationships between visuospatial ability, fluid ability and RPM across sex. Study 3 used meta-analytic techniques to synthesise research concerning sex differences on measures of inductive reasoning, considering the item stimulus and item type as potential moderators of this difference. Study 4 used exploratory and confirmatory structural equation modeling to examine the psychometric properties of the GRT2.

Results indicate that although the RPM tests are largely unidimensional, visuospatial ability is involved in performance. Furthermore, sex differences in raw scores and at the latent level were found, favouring males. Investigation of sex differences in inductive reasoning measures more broadly indicated that the figural format of these tests may contribute to the male advantage often identified; however, examination of the influence of the stimulus and type of question used in reasoning items in the GRT2 indicated that these do not meaningfully impact the latent construct measured.

# Declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I give consent to this copy of my thesis when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

I acknowledge that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Nicolette Waschl Date: 29/03/17

# Acknowledgements

Firstly, I would like to express my sincere gratitude to my principal supervisor, Professor Nick Burns; thank you for you patience, kindness and knowledge. I am grateful to have had a supervisor whom I always felt had my best interests at heart and whom I knew I was always able to approach with any concerns or questions I may have had. Thank you also to my co-supervisors Emeritus Professor Ted Nettelbeck and Dr. Irina Baetu, for their support during my PhD study. To Ted for always taking an interest in my work and providing invaluable feedback, and to Irina for the programming lessons and helpful chats.

My sincere thanks also goes to others who have enabled me to conduct this research; Dr. Simon Jackson, who provided data used in Papers 1 and 2, Stephen Kohl who provided data used in Paper 4, and Professor Andy Baker who was involved in providing data used in Paper 2. Also thank you to those authors who kindly provided me with unpublished information subsequently included in Paper 3. Thank you to Professor Jane Mathias and Associate Professor Siva Alagumalai for sharing with me their knowledge of meta-analysis and Rasch analysis.

I would like to thank Daniel McCluskey for his support during my time spent as a PhD candidate; you have been endlessly patient and understanding with my never-ending student status. Thank you also to my parents who have supported and encouraged me always, and to mum for taking the time to proof this document. Finally, thank you to my fellow PhD students and friends, particularly Erica, Michael and Jen, for all the coffee and beer (and wine and cheese).

# Chapter 1: Literature Review and Introduction

Intelligence testing has a long history in psychology, with the first intelligence test developed early in the 20th century. Since that time, many more cognitive ability tests have been created, and our understanding of intelligence theory has greatly improved. However, there remain many unknowns in the area. To this day, intelligence researchers are unable to agree on a single definition for intelligence, and our understanding of the processing requirements behind successful performance on cognitive ability tests remains incomplete.

The purpose of this thesis is to investigate the properties of one particular set of intelligence tests, the Raven's Progressive Matrices (RPM), and in so doing, to further understanding of what these tests measure, the relationship between fluid and visuospatial abilities, and sex differences in these abilities. Although the RPM were created nearly a century ago, there is continued disagreement over what it is that they actually measure. Furthermore, questions remain regarding the existence and implications of sex differences on these tests, and therefore the interpretation of scores.

The following review of the literature will consider the development and current state of research into the structure of cognitive abilities, with a specific focus on fluid and visuospatial abilities. Following this, research regarding which construct or constructs are measured by the RPM will be considered, and evidence for the claim that the test requires visuospatial ability in addition to fluid ability will be evaluated. Finally, sex differences in cognitive abilities and the implications of these differences for previous findings regarding the RPM and what the RPM tests measure will be considered, as well as the broader implications for other tests of fluid ability.

## Intelligence Theory

Throughout the history of intelligence research, many theories of the structure of cognitive abilities have been proposed; two main theories of note that have led to the development of psychometric intelligence theory today were proposed by Spearman (1923) and Thurstone (1938). In the early 20th century, Spearman proposed his two-factor theory of intelligence. Using what was then a relatively new technique, factor analysis, Spearman found evidence for the positive manifold, the fact that scores on different cognitive ability tests were positively correlated, meaning that those individuals who performed well on one test would typically perform well on others. Based on this, Spearman maintained that there was a general factor of cognitive ability that could be measured; a *g* factor. Other variance in test scores was considered specific variance associated with particular tests. As such, performance on cognitive ability tests was determined by both the *g* and *s* factors.

The main competition to Spearman's two-factor theory was Thurstone's (1938) theory of primary mental abilities. Thurstone maintained that intelligence was not a function of *g*, but rather consisted of several different, but inter-correlated abilities, similar to the broad factors we know now, such as memory, reasoning and visuospatial ability.

Intelligence theory today has evolved to incorporate elements of both Spearman and Thurstone's theories; typically intelligence is conceived as hierarchical, involving both broad abilities, which can be likened to Thurstone's primary mental abilities, and general and specific factors. Following is a discussion of current conceptualisations of the structure of cognitive abilities.

## Cattell-Horn-Carroll Theory

Currently, the most widely accepted and empirically verified framework for understanding the structure of intelligence is the Cattell-Horn-Carroll (CHC) model (Schneider & McGrew, 2012). This model is an amalgamation of Cattell and Horn's Gf-Gc theory (Cattell, 1941; J. L. Horn, 1965) and Carroll's (1993) Three-Stratum theory. Both of these theories are hierarchical models of intelligence; that is, they postulate specific abilities at the first stratum and broad group factors at the second stratum.

The two broad ability factors of Gf-Gc theory, fluid ability (Gf) and crystallized ability (Gc) were identified using second-order factor analysis. The Gf factor represented biologically influenced reasoning abilities, while the Gc factor represented information learnt from cultural exposure through both education and experience. Eventually, Gf-Gc theory evolved to include several more broad abilities, up to a total of eight additional abilities, when it was found that Gf and Gc alone did not satisfactorily represent relationships between other intellectual functions (e.g. visual, auditory and memory functions; J. L. Horn & Blankson, 2005).

Carroll's (1993) three-stratum theory was developed from an extensive factor analysis of ability test data collected over several decades. This theory was based on and to a large extent agreed with Cattell and Horn's Gf-Gc theory, but proposed three stratums instead of two; that is, it supported the existence of a general intelligence factor at the third stratum. Cattell and Horn both maintained that there was no general intelligence, and that any positive correlations observed between the second-order factors they identified were due to a statistical regularity caused by the difficulty in defining human action as only depending on one second-order ability (Hunt, 1999). Carroll, however, disagreed, maintaining that the correlation between second-order abilities did indicate a third-order general intelligence factor (Hunt, 1999). The existence or not of a general intelligence factor is an interesting debate in its own right, but it is not directly pertinent to the argument of this thesis and, as such, will not be discussed further.

Although there were some differences between the Gf-Gc theory and the three-stratum theory, such as differences in the number and type of broad abilities and the existence of a general intelligence factor, they can both be thought of as variations of the same hierarchical structure of cognitive abilities, and as such were first integrated in a published form by McGrew (1997), who presented a synthesised version of the two models, now called the CHC model. This original version of the model has since been expanded and modified, and now consists of over 70 narrow abilities, 16 broad abilities (see Table 1.1) and a general intelligence factor (Schneider & McGrew, 2012).

**Broad abilities under the CHC model.** The CHC broad abilities of particular concern to the present thesis are fluid reasoning and visuospatial ability. Fluid reasoning (Gf), involves solving unfamiliar problems for which existing knowledge cannot be used (Schneider & McGrew, 2012). Under the CHC model, Gf is conceptualised as involving three first stratum factors: induction, the ability to discover underlying rules or concepts and apply these to the problem; general sequential reasoning (deduction), the ability to apply known rules; and quantitative reasoning, the ability to reason using quantitative stimuli. Of main interest to the present thesis is the narrow ability of induction, which is the most representative of Gf (Carroll, 1993). Tests of induction require the individual first to discover the rule governing the relationship between the test stimuli and then either to state the rule or to apply it to the problem by selecting another exemplar to which the rule applies (Carroll, 1993).

Typically, induction is measured using series, analogies, classification and matrix tasks, with the RPM tests being one of the most characteristic measures commonly used.

Table 1.1

*Description of the sixteen broad abilities under the CHC model*

| Broad Ability | Abbreviation | Definition |
|---|---|---|
| Quantitative Knowledge | Gq | Knowledge related to mathematics |
| Reading & Writing | Grw | Knowledge and skills related to written language |
| Comprehension-Knowledge (Crystallised Ability) | Gc | Knowledge and skills valued by the individual’s culture |
| Domain Specific Knowledge | Gkn | Depth, breadth and mastery of specialised knowledge |
| Fluid Reasoning | Gf | Solving unfamiliar problems that cannot be solved by relying on previous knowledge only |
| Short-term Memory | Gsm | Encoding, maintaining and manipulating information in immediate awareness |
| Long-term Storage and Retrieval | Glr | Storing, consolidating and retrieving information over minutes, hours, days and years |
| Processing Speed | Gs | Speed of performance on simple repetitive cognitive tasks |
| Reaction & Decision Speed | Gt | Speed of making simple decisions when items are presented one at a time |
| Psychomotor Speed | Gps | Speed and fluidity of physical body movements |
| Visual Processing (Visuospatial Ability) | Gv | Making use of simulated mental imagery to solve problems |
| Auditory Processing | Ga | Detecting and processing meaningful non-verbal information in sound |
| Olfactory Abilities | Go | Detecting and processing meaningful information in odours |
| Tactile Abilities | Gh | Detecting and processing meaningful information in haptic sensations |
| Kinaesthetic Abilities | Gk | Detecting and processing meaningful information in proprioceptive sensations |
| Psychomotor Abilities | Gp | Performance of physical body motor movements with precision, co-ordination or strength |

*Note.* Definitions adapted from Schneider and McGrew (2012).

Gf is often thought to be at the core of what is generally meant when referring to general intelligence and is seen by some as equivalent to *g* (Gustafsson, 1984), although this is far from an agreed stance (Gignac, 2007). While the issue of the relationship between *g* and Gf are not a concern of the present thesis, it is important to keep this in mind when considering the importance and implications of what is known about Gf, and those tests that claim to measure it.

Visuospatial ability (Gv) is defined as the ability to deal with visual forms that are complex or difficult to perceive and manipulate (Schneider & McGrew, 2012). What Gv refers to is not simply the perception and encoding of visual stimuli, but the consequent mental operations performed on or with the stimuli, and hence it is far more complex than simple visual perception. There are various narrow abilities included under the Gv umbrella, such as visualization, spatial relations, closure flexibility, speed of closure and imagery. Of main concern to the present thesis are the narrow abilities of visualization and flexibility of closure. Visualization involves the ability to transform or manipulate objects and visual patterns and to imagine how they would appear under different conditions, without regard to the speed of these processes (Schneider & McGrew, 2012). Included under the narrow ability visualization is mental rotation, which tends to show very large sex differences in favour of males (Linn & Petersen, 1985; Voyer, Voyer, & Bryden, 1995) and because of this is often viewed as somewhat different to other visualization abilities. Flexibility of closure concerns the ability to identify a known stimulus embedded within a complex figure (Schneider & McGrew, 2012). These three narrow abilities demonstrate a close relationship with certain measures of induction and may play a role in performance on typical measures of inductive reasoning.

***The relationship between Gf and Gv.*** There are some uncertainties regarding the relationship between Gf, specifically the more narrow component inductive reasoning, and Gv. Historically, research into human cognitive abilities has struggled to differentiate reliably between Gf and Gv, often reporting a combined Gf/Gv factor (Crawford, 1991; Salthouse, Pink, & Tucker-Drob, 2008), or correlated residuals of the Gf and Gv factors (Reynolds, Hajovsky, Niileksela, & Keith, 2011).

A closer look at Carroll’s (1993) factor analysis shows that although separate Gf and Gv factors were found, certain tests of Gv also loaded onto the Gf factor, and vice versa. Although visualization was the narrow ability, of all the first-stratum abilities, which loaded most highly and consistently on Gv, it was also found to be the second most commonly loading ability on the Gf factor, and in some cases has been shown to load more highly on a Gf factor than a Gv factor (Schneider & McGrew, 2012). Thus, although visualization is clearly visuospatial in nature, and hence can be considered a Gv ability, its high and consistent loadings on the Gf factor may be considered problematic when attempting to separate Gf from Gv.

Similarly, inductive reasoning, the most representative and consistently high loading narrow ability on Gf, showed moderate loadings on the Gv factor in Carroll’s (1993) analysis. It is unclear whether these results occurred because visualization involves some kind of inductive reasoning, because inductive reasoning tests may often require Gv abilities, or because these two abilities involve some other similar process. Thus, the relationship between these abilities is still not well understood, and confusion regarding this relationship can make the interpretation of what ability tests measure all the more difficult.

One of the problems regarding the separation of Gf and Gv may lie in the fact that many tests of Gf, and particularly tests of inductive reasoning, use visual stimuli.

While Gv abilities are associated with the manipulation of visual forms as such, it is unclear the extent to which they may also be involved in tests that utilize a visual format but are hypothesized to measure other abilities, such as the abstract figures commonly used to measure inductive reasoning. Additionally, ability tests are often not factorially unidimensional (Carroll, 1993). Hence, if a factorially unidimensional measure of induction cannot be found, or if a factorially unidimensional measure of visualization cannot be found, then it will continue to be difficult to separate the Gf and Gv factors.

Compounding the problem of distinguishing between Gv and Gf is the fact that the Gc/Gf separation is often confounded with a verbal/non-verbal (visuospatial) distinction (Major, Johnson, & Deary, 2012). One of the pitfalls of the factor analytic methodology lies in the fact that it is dependent on the tests included in the analysis. Hence, a factor analysis of any battery of tests will only find those factors adequately represented in that battery, and the factors found will only contain tests of a certain nature if they are included in the battery. If all tests representing a Gc factor are of a verbal nature, while all tests representing a Gf factor are of a visuospatial nature, the Gc factor will likely represent, to an extent, verbal ability, while the Gf factor will likely represent, to an extent, Gv, even though these are not necessarily characteristics of Gc and Gf. However, without a sufficient number of Gc measures that do not utilize a verbal format, and of Gf that do not use a visual format, it is difficult to determine what the relationship is between the Gf-Gc dimension and the visual-verbal dimension. In addition to this, the interpretation of the results of factor analysis presents certain issues. Factor analysis is able to determine the existence of common variance, but not necessarily what it is that the common variance represents; the interpretation of the meaning of that variance can be influenced by the assumptions held by those interpreting it.

As extensively discussed by Wilhelm (2005), an alternative conceptualisation of the distinction between induction, deduction and quantitative reasoning is a distinction between the type of stimuli used in the tests, referred to as the content facet. Carroll's (1993) factor analysis, which is the basis of the existence of the three first-stratum Gf abilities, demonstrated that induction tended to be measured using figural stimuli, deduction using verbal stimuli, and quantitative reasoning using numerical stimuli. This brings to attention one of the main issues regarding the CHC taxonomy and what it says about reasoning ability; namely, what, if any, the role of the content facet is. Wilhelm (2005) found that induction and deduction were factorially identical and that a distinction between the content facets better explained the structure of reasoning ability than did a distinction between inductive and deductive reasoning.

Yet another confounding factor in the relationship between Gf and Gv concerns the notion of the use of different strategies for obtaining the correct solutions. As argued by Wilhelm (2005) and Schneider and McGrew (2012), many Gv tests can be solved using either a visuospatial or an analytic strategy, which may help to explain the confusion regarding the relationship between Gv and Gf; if some individuals use visuospatial strategies, while others use analytical reasoning strategies, then it is conceivable that these tests could load on two factors: Gv and Gf. Furthermore, as the authors argue, one mark of success on these tests may be the ability to switch between strategies, and to use the most appropriate strategy for each item, which would differ depending on the particular idiosyncratic features of that item. This type of strategy switching ability may be related to Gf. Unfortunately, given the confounding issues of stimulus type and strategy use in tests of Gf, it is hard to determine what the relationship is between Gf and Gv.

## Other Theoretical Frameworks

Although the CHC model is the most accepted taxonomy of cognitive abilities, this does not necessarily mean that it provides the definitive answer to how cognitive abilities are structured. Alternative models of the structure of intelligence may be able to shed light on the relationship between Gf and Gv through different conceptualisations of how broad abilities relate to one another and what role Gv plays in the structure of cognitive abilities. As such, although a large part of this thesis is framed under the CHC model of cognitive abilities, two alternative models will be briefly considered due to the different perspectives they are able to provide; the Berlin Model of Intelligence Structure (BIS) and the Verbal-Perceptual-Image Rotation model (VPR).

The BIS is a faceted model of intelligence, conceived by Jäger (1982). The BIS distinguishes between operation and content components of ability tests, and therefore postulates a 4x3 (operations x contents) matrix to represent the structure of cognitive abilities. The four operations are processing speed, memory, creativity and processing capacity (or reasoning); and the three contents are figural, verbal and numerical. Although the operation specific variance in measures of intelligence has been found to be stronger than content specific variance, a model accounting for both of these types of variance has shown a better fit to the data, supporting the validity of both dimensions (Bucik & Neubauer, 1996). Processing capacity can be considered the equivalent of Gf under the CHC model, and like Gf has been found to be closely related to the *g* factor (Bucik & Neubauer, 1996). Therefore, according to the BIS, Gf can be measured with figural, verbal or numerical content. Note that this maps nicely on to the distinction between induction, deduction and quantitative reasoning previously discussed. Given this, the BIS is not necessarily in conflict with the CHC model, although it highlights a different way of conceptualising different types of reasoning, and the relationship between Gf and Gv. Accordingly, more recently, there has been discussion of the relevance of the BIS theory to CHC theory, and how BIS might be incorporated into the CHC model (Schneider & McGrew, 2012).

The Verbal-Perceptual-Image Rotation (VPR) model (Johnson & Bouchard, 2005a, 2005b), based on Vernon's (1964) verbal-perceptual model, is another hierarchical model with a general intelligence factor at the highest stratum, group factors (verbal, perceptual and image rotation) and specific factors on lower strata. This model highlights the importance of Gv measures, by placing them at the same level as Gc measures. Because this model considers Gf to be functionally equivalent to *g*, tests conceptualised as measures of Gf under the CHC model by necessity must load on an alternate factor. In Johnson and Bouchard's (2005b) study they found that the two measures of Gf used (both inductive reasoning) showed primary loadings on the spatial factor (perceptual), and secondary loadings on either the number factor or scholastic (verbal) factor. Therefore, under this model, tests considered to measure Gf under the CHC model are actually determined to be measures of spatial ability, loading onto the same factor as measures of Gv. This provides another way of conceptualising Gf abilities, and suggests that they may be more similar to Gv abilities than to Gc or verbal abilities. Although the CHC model does allow groupings of the broad abilities according to conceptual or functional groups, the Gf and Gv factors do not tend to be related to one another. For example, in terms of conceptual groupings, Gv may be grouped with other sensory abilities, such as Ga, Go, Gv and Gh, while Gf may be grouped with other so-called domain-independent general capacities (Schneider & McGrew, 2012). Thus, the VPR model's alternative conceptualisation of the structure of intelligence, which permits a close relationship between inductive reasoning and visuospatial ability, may be useful in understanding the relationship between Gv and Gf.

## The Raven's Progressive Matrices

Raven's Progressive Matrices (RPM) are among the most popular ability tests worldwide. They are the second most commonly used group of psychological tests in the world and have attracted a large amount of research (Oakland, 1995; Oakland, Douglas, & Kane, 2016). These tests were designed in the 1930s by J. C. Raven who believed that full-length intelligence tests were difficult to administer and results difficult to interpret if the examinee was illiterate or unable to follow instructions (Raven, 2008). His aim was therefore to develop a test that was easier to administer, theory-based and directly interpretable, unlike full-length intelligence test subscales (Raven, 2008). The RPM were originally designed to measure one of the two main components of general intelligence identified by Spearman (1927), eductive ability, or "meaning-making". Nowadays, this set of tests is considered among the best measures of Gf (McGrew & Flanagan, 1998), although often also used as a measure of *g* due to Gf often being conceptualised as equivalent to *g* (Gustafsson, 1984; Kvist & Gustafsson, 2008; Major et al., 2012). The RPM tests have been very popular as a result of the perceived relatively small influence of language and culture on performance, and because of their ease of administration to large groups of people.

The RPM tests consist of either 2x2 or 3x3 matrices of figural patterns or designs with one part missing. The task is to determine from eight possible solutions, which solution correctly completes the pattern. There are now at least three full-length versions of this test for individuals of differing levels of ability, and several short forms. The Coloured Progressive Matrices (CPM) was designed for young children, the elderly, and individuals with an intellectual disability. The Standard Progressive Matrices (SPM) was designed for use in the general population. This test consists of 60 progressively more difficult items, divided into five sets of 12 items (Sets A through E). The Advanced Progressive Matrices (APM) was designed for adolescents and adults of above-average ability. This test consists of 48 items, divided into Set I (12 items) and Set II (36 items). Set I of this test samples the full range of difficulty in the SPM, and is often used for practice before completing Set II. Typically, APM total scores are calculated from the items of Set II only.

This literature review will only consider findings related to the SPM and APM, because the CPM has typically been used for very young or intellectually disabled individuals, and may have slightly different properties. A large portion of the research into the dimensionality of the RPM and its association with Gv has been conducted with the APM, most likely because of its suitability for administration to a university population, with which most of the associated research has been conducted. However, where applicable, findings pertaining to the SPM will also be discussed.

## Validity and Reliability of the RPM

The RPM tests have shown good validity and reliability. These tests have shown good convergent validity with other measures of inductive reasoning (Schweizer, Goldhammer, Rauch, & Moosbrugger, 2007), as well as good internal consistency (Arthur & Day, 1994; Paul, 1985) and test-retest reliability (Bors & Forrin, 1995). The tests have been found to measure much the same thing across socio-economic and cultural groups (Raven, 2008). For example, an analysis of the 1979 British standardisation data showed that correlations between item difficulties ranged from .97 to .99 in children from eight different socio-economic groups (Raven, 2008). Similarly, analysis of the US standardisation data demonstrated a correlation ranging from .97 to 1.00 between item difficulties among five separate ethnic groups (Raven, 2008). However, whether this finding applying to a US sample can be extrapolated to other cultural groups not residing within the US is questionable (Owen, 1992).

Although the RPM tests were created based on Spearman's theory of *g*, and are widely considered a good measure of Gf, there have been various claims that these tests also require Gv for successful performance (e.g. Burke, 1958; DeShon, Chan & Weissbein, 1995) and, indeed, there are questions surrounding the discriminant validity between scores on the RPM tests and measures of Gv (Mackintosh & Bennett, 2005; Schweizer et al., 2007). Research has followed many different paths in order to investigate the role of Gv in RPM performance. Attempts have been made at creating taxonomies to describe the rules involved in solving items and corresponding analyses of these taxonomies have been conducted. Investigations of the correlations among the RPM and other measures of reasoning and Gv have also appeared in the literature. However, it is still unclear to what extent Gv plays a role in RPM performance, if at all. Following is a discussion of the evidence both for and against the role of Gv in RPM performance, as well as a consideration of the issues confronted in the investigation of this question.

## The Dimensionality Problem

One of the first steps in determining if Gv is involved in RPM performance has been to examine the dimensionality of the tests; that is, do all test items measure one construct (unidimensionality) or do some items measure an additional construct (multidimensionality)? However, determining how unidimensionality should be established can be a problematic issue. It can be argued that no measure can ever be truly unidimensional, because it may be unrealistic to expect that any human task could involve only one ability. However, to what extent can a test be multidimensional before this presents a problem to the assumption of unidimensionality and the interpretation of scores? This is a difficult question to answer. Two issues pertaining to this question are an understanding of (i) psychometric and (ii) psychological unidimensionality, and the distinction between these two concepts.

First, psychometric unidimensionality concerns the notion that all non-random variance in a dataset can be accounted for by a single latent construct, or dimension of difficulty (Sick, 2010). In testing for psychometric unidimensionality, factor analysis has been a popular method. Factor analysis falls under the theoretical orientation of classical test theory (CTT), which has traditionally been used in the construction and validation of psychological tests. However, there are various problems with the use of factor analysis to answer the question of dimensionality in the RPM; factor analysis was not designed for use with binary data such as the correct-incorrect responses obtained from the RPM, and when used with these type of data can produce artifactual factors based on difficulty (Gibson, 1960; Hattie, 1985). Newer methods of factor analysis have been developed that may circumvent this problem (e.g. WLSMV estimation; Muthén, du Toit, & Spisic, 1997) but these are not yet widely used.

Another method of assessing psychometric dimensionality involves item response theory (IRT), which provides an alternative method of conceptualising and analysing test validity. In contrast to CTT, IRT was created for use with, and is therefore appropriate for, binary data. Certain IRT models, such as the Rasch model, assume unidimensionality. As such, item fit and residual variance can be used to investigate test dimensionality. The combined use of CTT and IRT approaches when assessing dimensionality can be complementary.

However, even if satisfactory psychometric unidimensionality is established, this still leaves the problem of the interpretation of what this unidimensionality substantively means; that is, does this finding also signify psychological unidimensionality, or more simply put, that the test measures only one ability? Contrary to intuition, psychometric unidimensionality does not necessarily mean that the test measures only one construct but, rather, that every item measures the same thing. Therefore, if every item measured a combination of two abilities, the test could conceivably be psychometrically unidimensional, but also measure more than one latent construct. In short, psychometric unidimensionality may be a necessary, but not sufficient, requirement for psychological unidimensionality. Because of this, while an analysis of unidimensionality can tell us something about what the test measures, we must also look towards other evidence from different research paradigms to really answer the question of the extent of involvement of Gv in the RPM. Therefore, findings from other research paradigms must be considered in conjunction with findings regarding the dimensionality of the RPM.

**Proposed taxonomies of solution strategies.** One technique that has been applied in assessing the extent of the involvement of Gv in RPM performance is consideration of the types of rules individuals use to solve RPM items, or taxonomies of solution strategies. The most comprehensive of these have been developed with regard to the APM, with little work of this kind available on the SPM. Among the first attempts to classify the solution strategies required of the problems in the APM was

Hunt (1974). Hunt’s two broad categories of strategies for solving Set I of the APM were formed by inspection of the items and hypothesised approaches to obtaining the correct solution. Hunt’s work resulted in a computational model of problem solving that could use either a Gestalt (perceptual) strategy or an analytic strategy to solve items. Although Hunt’s work is interesting, there has been a paucity of studies regarding its validity.

Carpenter, Just, and Shell (1990) argued that the items of Set II of the APM cannot be solved using perceptual strategies like those that can be used to solve Set I. Carpenter et al.’s (1990) theoretical model of the processes involved in solving APM items was based on an inspection of item characteristics, consideration of the performance of average and above-average scorers, verbal reports by test-takers, and patterns of errors and eye fixations during test performance. According to this model there are five analytic rules that can be used to describe the relationship between the correct answer and the rest of the matrix. In order to solve many of the problems, more than one rule, or more than one instance of a particular rule, is needed, but only analytic rules are required (see Table 1.2).

Carpenter et al.’s (1990) taxonomy has proved a useful and popular method of analysing and understanding the types of problem solving that may be involved in the APM. However, DeShon et al. (1995) argued that it was inadequate in that there was no consideration of visuospatial processes that may aid in solving the items. DeShon et al. maintained that visuospatial processes were involved in APM performance and expanded Carpenter et al.’s rules to 10 different solution rules; four of which involved verbal-analytic processes, and six of which involved visuospatial processes (Table 1.3).

Table 1.2

*Solution strategies from Carpenter et al. (1990)*

| Rule | Description |
|---|---|
| Constant in a row | The same value occurs throughout a row, but changes down a column |
| Quantitative pairwise progression | A quantitative increment or decrement occurs between adjacent entries in an attribute such as size, position, or number |
| Figure addition or subtraction | A figure from one column is added to (juxtaposed or superimposed) or subtracted from another figure to produce the third |
| Distribution of three values | Three values from a categorical attribute (such as figure type) are distributed through a row |
| Distribution of two values | Two values from a categorical attribute are distributed through a row and the third value is null |

Table 1.3

*Solution strategies from DeShon et al. (1995)*

| Rule | Description |
|---|---|
| Visual | |
| Superimposition | The perceptual representation of an object is mapped onto a second object by placing the borders of the two objects in correspondence. The new image is composed of the overlapping borders of the two objects and the sum of each object's unique features |
| Superimposition with cancellation | Special case of the superimposition rule. Objects are placed in correspondence by overlapping their respective borders and features that overlap cancel each other out |
| Object addition/subtraction | Visually combining two objects into a whole. The objects are not superimposed, but placed next to one another on the basis of a common border. |
| Movement | Objects move incrementally from frame to frame with respect to the stable background |
| Rotation | The degree of rotation required to bring objects into correspondence from time 1 to time 2 must be equivalent to the degree of further rotation required to bring objects from time 2 to time 3 into correspondence |
| Mental Transformation | Performing an operation on an object in the third entry that is specified by the second entry |
| Analytic | |
| Constant in a row | A feature is identical across the row, but changes down the column |
| Quantitative pairwise progression | A quantitative increase or decrease in the number of features from square 1 to square 2, and square 2 to square 3 |
| Distribution of 3 values | Three values from a categorical attribute (e.g. shapes) are distributed through a row or column |
| Distribution of 2 values | Two values from a categorical attribute are distributed through a row and the third value is null |

*Note:* Two other rules also exist in DeShon et al.'s (1995) rule system, although the authors do not formally describe them. Superimposition with conditional placement is a form of superimposition, where the elements of the figure are placed relevant to another feature of the figure (e.g. a circle is placed on top of a cross but a square is placed behind a cross). This rule occurs a total of three times and can be thought of as a special case of superimposition. The expansion rule is a form of the movement rule, where the figure is perceived to expand. This rule occurs once in the APM.

However, the literature has been inconclusive regarding the validity of these taxonomies in explaining any apparent multidimensionality in the APM. Research has examined whether the different rules in these taxonomies may represent different latent performance factors. These different factors have typically been conceptualised as four factors in the case of Carpenter et al. (1990), based on the 4 different solution rules (excluding constant in a row, because it always occurs in conjunction with one of the other rules). In the case of DeShon et al. (1995), two factors based on the distinction between the analytic-visual classifications have been proposed. Abad, Colom, Rebollo, and Escorial (2004) compared the fit of DeShon et al.'s two-factor model and a one-factor model. They found that the one-factor model, where each item was conceived as measuring the same construct, most likely simply a single reasoning factor, provided a better fit. This is not surprising, in that DeShon et al. themselves actually found that a two-factor model based on their taxonomy did not show a better fit than a one-factor model. However, they continued to argue for the validity of their two factors by virtue of the evidence from their experimental manipulation of concurrent verbalization and their inspection of the items. Similarly, using factor analysis, Vigneau and Bors (2008) found little support for a latent variable conceptualisation of DeShon et al.'s verbal-analytic distinction or for Carpenter et al.'s four different solution rules.

On the other hand, some research does support DeShon et al.'s (1995) analytic-visual distinction. For example, Borst and Kosslyn (2010) investigated the correlations between several Gv measures, and the analytic and visuospatial items as classified by DeShon et al., finding that spatial imagery and visualization were significantly correlated with visual but not analytic items.

Although taxonomic work has not been as well advanced with the SPM as with the APM, the solution taxonomies constructed with regard to the APM may be useful in also understanding how SPM items are solved. Carpenter et al. (1990) and DeShon et al.'s (1995) rules can be applied to many of the later items in the SPM, although the earlier items are probably too simple for the application of these rules, tending to involve simpler perceptual processes.

**Factor Analysis**. There has been a large amount of research conducted concerning the factor structure of the APM. One seminal paper is that of Dillon, Pohlmann, and Lohman (1981), which found that scores on the APM were represented by two orthogonal factors; addition/subtraction and pattern progression. Addition/subtraction was described as the ability to add and subtract patterns while pattern progression was described as the ability to perceive the progression of a pattern, and hence both can be thought of as types of visuospatial strategy. Since publication of this paper, there have been many attempts to validate the existence of these two factors. For example, Alderton and Larson (1990) attempted to replicate Dillon et al.'s work, but could not confirm the factor structure found by the original authors. Although a two-factor solution was investigated, it was found to be uninterpretable and unstable across samples. Hence, they concluded that a one-factor model, most likely representing a sole Gf factor, best represented scores on the APM. Similarly, other studies (Abad et al., 2004; Arthur & Day, 1994; Arthur & Woehr, 1993; Bors & Stokes, 1998) have compared the fit of Dillon et al.'s two-factor model to other models including a one-factor model and correlated two-factor models. All of these studies found that either a one-factor or correlated two-factor model showed a better fit than the model based on Dillon et al.'s work. Where a correlated two-factor model was found to fit well, the two factors were very highly correlated and hence it was again concluded that a one-factor model best represented the data.

The lack of confirmatory evidence regarding the factors proposed by Dillon et al. (1981) is hardly surprising. In order to obtain a two-factor solution, an orthogonal rotation was performed. This method of rotation assumes uncorrelated factors; however, cognitive abilities have consistently been found to be positively correlated. Therefore, to assume that two different ability factors measured by the APM are uncorrelated makes little theoretical sense. Further, only 15 of the 36 items of the APM were found to be relatively pure measures of the two factors, indicating that this classification system does not apply to many of the items in the test.

Dillon et al.'s (1981) factor structure has also been compared to the solution taxonomies of Carpenter et al. (1990) and DeShon et al. (1995). Abad et al. (2004) found that neither Dillon et al. nor DeShon et al.'s models fit better than a one-factor model. Vigneau and Bors (2008) compared the fit of several models, based on Carpenter et al., DeShon et al., Dillon et al., and item skewness, respectively. It was found that the model based on item skewness, representing a difficulty or item position factor, provided the best fit to the data, although fit statistics were quite similar across all models. This indicates that while the taxonomies of solution strategies have been quite popular and well used, and may explain the rules individuals apply in solving APM items, none of them satisfactorily describes the structure of the APM. There are two ways in which the superior fit of the item skewness model can be interpreted; either this was the result of the method of factor analysis used, or there is some substantive difference between the items at the beginning and end of the test. Maximum likelihood confirmatory factor analysis (CFA) was used to test the fit of the five models. The use of the maximum likelihood method can produce erroneous results when the data have a small number of categories and do not meet the multivariate normality assumption (Mindrila, 2010), both of which are the case with data derived from the APM.

However, results from Vigneau and Bors' (2005) study suggested that a qualitative difference may exist between the items at the beginning and end of the test. This idea is further supported by the solution taxonomies, for which the pattern of solution rules tends to involve certain rules used at the beginning, while other rules are used for the items towards the end of the test. Furthermore, this idea brings up the issue of the relevance of difficulty factors. If it is impossible to separate Gv from item difficulty, then it is difficult to find the presence of a visuospatial factor in the RPM. It is possible that the more consistent findings of multiple factors in the SPM than in the APM may be due to the easier format of the SPM. It may be that when test items are more difficult, there is a greater tendency to use analytic rather than spatial strategies (Kyllonen, Lohman, & Snow, 1984). Interestingly, findings by Schweizer, Schreiner, and Gold (2009) indicate that, while a two-factor model can demonstrate a good fit to APM data, this two-factor model may be best represented by a learning effect. This finding also needs to be kept in mind when considering whether a lack of fit for a one-factor model necessarily means that the APM measures at least two distinct constructs.

A large number of other studies have investigated the factor structure of the APM without explicitly testing any specific taxonomy. Some of these studies have found the best fit with a single-factor model (Arthur, Tubre, Paul, & Sanchez-Ku, 1999; Chiesi, Ciancaleoni, Galli, Morsanyi, & Primi, 2012; Chiesi, Ciancaleoni, Galli, & Primi, 2012), while others have found more than one factor (Bors & Vigneau, 2001; Lim, 1994), and yet others have found conflicting evidence (Vigneau & Bors, 2005). The study by Bors and Vigneau (2001) investigated the effect of practice on APM performance. Although this study was not explicitly designed to assess the factor structure of the APM, some interesting findings were produced. Factor analysis indicated only a borderline adequate fit for a one-factor model across three testing sessions, raising the possibility that the APM does in fact measure more than one factor. However, the more interesting finding from this study was that practice did not improve the fit of the one-factor model. Bors and Vigneau (2001) argued that, if the APM did indeed measure only one latent factor, the fit of the one-factor model would be expected to improve across testing sessions, because practice decreases random error variance in measurement. Hence, the fact that the fit of the model did not improve indicates that it may not be the most appropriate model for the structure of the APM.

Finally, results of factor analyses by Vigneau and Bors (2005) have been inconclusive with regard to the number of factors present in the APM. Using principal components analysis, a one-factor structure was supported, but it explained a smaller amount of variance than desirable. Nonlinear factor analysis, however, indicated the existence of two factors. Given the disadvantages of principal components analysis as already highlighted, it can be argued that this work indicates the existence of more than one factor in the APM. Furthermore, this study has served to highlight how the use of different factor analytic methods can produce different results, and hence the importance of using the most appropriate method available.

Turning to the SPM, factor analytic results show a similar pattern, although there is perhaps more support for multiple factors in this version of the test. Lynn, Allik, and Irwing (2004) used exploratory factor analysis, followed by confirmatory factor analysis, to determine a three-factor structure for the SPM. These factors were interpreted as Gestalt, verbal-analytic and visuospatial. Interestingly, their interpretation of these factors is somewhat contradictory. While DeShon et al. (1995) tended to classify those items observed to require the addition/subtraction rule as visuospatial, with Mackintosh and Bennett (2005) following suit, Lynn et al. argued that this type of item is verbal-analytic. However, regardless of the labels assigned to the factors, later work in different samples has supported this three-factor structure (Bakhiet, Haseeb, Seddieg, Cheng, & Lynn, 2015; Grigoriev & Lynn, 2014). Nonetheless, to this point in time, this work is yet to be replicated by an independent group not associated with Lynn.

On the other side of the argument, factor analytic studies from African populations have used both principal components analysis and confirmatory factor analysis to investigate the factor structure of the SPM, with results supporting unidimensionality (Abdel-Khalek, 1988; Al-Shahomee, 2012). Other work has found that the first factor obtained from the SPM data is so dominant as to render any interpretation of the subsequent factors meaningless (Cikrikci-Demirtasli, 2000). Likewise, using confirmatory factor analysis of subscale scores, Arce-Ferrer and Guzmán (2009) found support for a one-factor structure. However, while these authors argued for the use of subscale scores in order to minimize violations of the linearity assumption that can result when subjecting binary scores to a factor analysis, this probably influenced the results found. By nature, the use of subscale scores would result in the loss of some of the individual variance of items, therefore potentially causing the appearance of a single factor.

**Item Response Theory Analysis**. Studies using IRT methods to investigate the dimensionality of the RPM have also produced conflicting results but have tended to support multidimensionality of both the APM and SPM. Vigneau and Bors (2005) found, despite support for a one-factor CFA model, that Rasch analysis indicated violations of unidimensionality in the APM. In fact, this analysis indicated that approximately half of the items would need to be deleted before homogeneity, and therefore unidimensionality, could be achieved. Interestingly, findings indicated that subsets of items created according to item difficulty were unidimensional, pointing towards a qualitative change in addition to a quantitative change across test items.

In terms of the SPM, IRT analyses have both supported (Van der Elst et al., 2013) and refuted (Kubinger, Formann, & Farkas, 1991; van der Ven & Ellis, 2000) unidimensionality. Van der Elst et al. (2013) first determined unidimensionality by using Monte Carlo simulations to compare the second eigenvalue of their data with the second eigenvalue of data randomly generated by the IRT model, finding no significant difference between the two, and therefore providing evidence of unidimensionality. They applied both the one parameter (1PL) and two parameter (2PL) IRT models to Sets B, C and D of the SPM. The 1PL and 2PL are different forms of the IRT model; the 1PL model considers the probability of a correct response to an item given the individual's ability and item's difficulty, but constrains equal item discrimination (how well an item discriminates between individuals with different levels of ability) and guessing. The 2PL model considers the probability of a correct response given ability, difficulty and discrimination, holding only guessing constant. From their analysis, Van der Elst et al. found that eight items (22%) did not fit the 1PL model, while three items (8%) did not fit the 2PL model. While the results for the 1PL model are worrying, those for the 2PL model are acceptable. However, given that the analysis only included three of the five sets of the SPM (B, C and D), this would likely have influenced dimensionality results. This can be seen from the factor analysis of Lynn et al. (2004), who found that items from Sets A and E were those that most loaded on their gestalt continuation and verbal-analytic reasoning factors respectively, but items from Sets B, C and D tended to load onto the visuospatial factor. In short, those three sets may be more similar than the other two.

Moreover, there is research that suggests both the SPM as a whole, and the individual item sets, may not be unidimensional. Van der Ven and Ellis (2000), using a Rasch model, looked at the dimensionality of each subset of the SPM. Their results indicated that sets B and E were not unidimensional, and that set C may not be either. Similarly, Kubinger et al. (1991) found that the 1PL model did not fit SPM data, and suggested that this means that either the 2PL or 3PL IRT model is necessary, or that the test is not unidimensional. However, given the strong learning effects they found in their data, they argued that it may be more likely that this was responsible for their findings.

**Other Research Paradigms**. The evidence for multidimensionality, and therefore that the RPM tests involve another ability in addition to inductive reasoning, is far from conclusive in the factor analytic and IRT literature. However, other methods used to investigate the relationship of Gv to RPM performance do indicate that Gv may play a role. These have included experimental methods, correlational methods, and brain imaging. Results from these other paradigms suggest that the investigation of the importance of the role of Gv in APM performance should not yet be rejected.

In terms of experimental work, DeShon et al. (1995) used a verbal overshadowing effect, a phenomenon that occurs when tasks requiring visuospatial processing are negatively affected by concurrent verbal processing, to investigate the viability of their verbal-analytic and visual strategy item classification of the APM. It was predicted that those participants required to concurrently verbalize their problem solving strategies would perform worse on items classified as requiring a visuospatial strategy than would those participants who completed the test according to standard procedure. This was what occurred; those participants in the concurrent verbalisation condition exhibited worse performance than did those in the control condition on 57% of the visuospatial items, but performance was equivalent on the analytic items. This suggested that verbal overshadowing was functioning to inhibit visuospatial processing, thereby negatively affecting performance on those items requiring this type of processing. Research by Borst and Kosslyn (2010) supported this, finding that measures of visualization and spatial mental imagery correlated significantly with total scores on the visual APM items, but not the verbal-analytic items. Further, a study by Prabhakaran, Smith, Desmond, Glover, and Gabrieli (1997) provided brain imaging evidence of a distinction between these two types of items. How these items were classified was somewhat different to DeShon et al.'s (1995) classification; they classified Carpenter et al.'s (1990) pairwise progression rule as visuospatial and the other three rules as analytic, whereas DeShon et al. classified addition/subtraction as visuospatial and the other three rules as analytic. Based on this distinction, however, they did find that although there was some bilateral activation for both item types, there was more left hemisphere activation for analytic items but more right hemisphere activation for visuospatial items.

Stephenson and Halpern (2013) investigated the influence of different types of working memory training on performance on four different measures of Gf, including the APM. There were five different conditions; a control condition, a dual n-back condition (where the working memory task included both auditory and visual stimuli), a visual n-back condition, an auditory n-back condition and a visual short-term memory training condition. Regarding the difference between pretest and posttest performance on the APM, all conditions apart from the auditory n-back condition showed greater improvement than the control condition, while the auditory n-back condition was not significantly different from any other condition. This provides some limited support for the notion that visuospatial processes play a role in APM performance. Similarly, Colom, Román, et al. (2013) found that working memory training including spatial and auditory components improved performance on the APM and the Differential Aptitude Tests – Abstract Reasoning subtest (DAT-AR; another Gf measure that utilizes visual stimuli) much more than it did on the Primary Mental Abilities – Reasoning subtest (PMA-R), a measure of inductive reasoning consisting of letter stimuli. Again, although the differences here were not significant, there was a weak trend favouring the role of visuospatial processes in APM performance.

An interesting finding emerged from a study designed to investigate the neural correlates of several different abilities, including induction and visualization. In this study, individuals were separated into higher and lower Gf groups according to their scores on the APM. It was found that those individuals in the lower Gf group made significantly more errors on visualization and spatial relationships, but not on their custom-made measure of induction, indicating that Gv processes may be involved in APM performance (Ebisch et al., 2012). Interestingly, this custom-made measure of induction also utilized figural stimuli, but used an analogical format of A:B::C:?. This indicates that it may not be the figural format of the RPM *per se* which involves Gv, but the operations that are required to be performed on these figural stimuli.

Colom, Stein, et al. (2013) administered several tests of inductive reasoning (APM, DAT-AR and PMA-R), as well as tests of Gv (DAT – Spatial Relations, PMA – Space, Rotation of Solid Figures) and Gc (DAT – Verbal Reasoning, PMA – Verbal, DAT – Numerical Reasoning) and investigated brain activation during performance on these measures. Specifically, the interaction between hippocampal activity and these ability measures was investigated. Colom, Stein, et al. found that, when interactions with sex were considered, DAT-AR and APM performance were related to different hippocampal areas: the DAT-AR to the left posterior hippocampus and the APM (and Rotation of Solid Figures) to the right posterior hippocampus. This provides additional support for the notion that scores on the APM may involve Gv, and to a greater extent than other Gf measures. While the DAT-AR is another test of inductive reasoning involving visual stimuli, the stimuli used in this test are less spatially complex than those used in the APM. Hence, it is possible that the complexity of the figural stimulus used in the APM may differentiate this test from other figural matrices tests. Given a common research finding that males perform better than females on tests of spatial visualization (Hunt, 2011), it is worth noting that other research does indeed suggest that there is something specific to the APM that may bias the test in favour of males. Arendasy and Sommer (2012) found that, out of several different figural matrices tests, the APM showed the most pronounced bias in favour of males and suggest that certain item design features, particularly element salience, may be responsible for this. The DAT-AR was designed specifically to avoid the issue of visual discrimination influencing performance, with the item stimuli described as “large and clear” and differences between diagrams as “obvious” (Kettner, Seashore, & Wesman, 1966).

Studies have also investigated the relationship between autism and autistic traits and performance on the RPM tests. Individuals with autism typically perform better on visually based tests such as the RPM. It is suggested that this is because of their superior ability to perform tasks involving disembedding, or separating a part of a stimulus from the whole (Fugard, Stewart, & Stenning, 2011). Fugard et al. (2011) administered a scale of autistic-like tendencies to undergraduate students and investigated how these scores related to performance on the APM. It was found that higher scorers (more autistic-like traits) were more accurate on items that could be classified as visual. The authors of this paper explored the possibility that the visual items of the APM were simply easier than the other items, but found no difference in the overall accuracy for visual versus analytic items. This suggests it is a characteristic peculiar to the visual items that is responsible for the relationship with autistic-like traits. Other research supports the notion that flexibility of closure, or disembedding ability, is involved in RPM performance. Several research studies have provided evidence that element salience of RPM items, operationalised as the ease of identification of specific elements, is an important difficulty factor in performance (Meo, Roberts, & Marucci, 2007; M. J. Roberts, Welfare, Livermore, & Theadom, 2000).

However, other evidence suggests that while the RPM tests may involve Gv to an extent, this is only minimal, and Gv has little influence on performance. Schweizer et al. (2007) investigated the convergent and discriminant validity of the APM. It was expected that the APM would demonstrate a considerable relationship with another measure of inductive reasoning, specifically W. Horn’s (1983) reasoning scale. This scale was chosen because the stimuli used are numbers and letters, as opposed to figures, and therefore the use of this measure would avoid any spurious correlations caused by the use of similar stimuli. In investigating discriminant validity, it was predicted that the APM should not correlate too highly with measures of other abilities, specifically W. Horn’s measures of visualization, mental rotation and speed of closure. The authors investigated three structural equation modeling (SEM) models: a correlation model and two prediction models. The correlation model demonstrated that the latent variable associated with APM scores correlated most highly with inductive reasoning. Although the latent APM variable was also significantly correlated with Gv, once the variance accounted for by inductive reasoning was removed from this correlation there was only a negligible correlation left. In the prediction models, reasoning was the only significant predictor and Gv did not improve prediction of the APM latent variable. Taken as a whole, the results of this study suggest that spatial ability does relate to APM performance; however, this relationship is not substantially important in APM performance, and the test can be considered a measure of reasoning. Nonetheless, one potential issue with this study was the use of item parcels in the analysis. The use of item parcels involves combining the scores on different items in order to decrease the number of data points, as well as to avoid the problems associated with using binary data. For example, if a test consists of 20 items, four item parcels could be created, consisting of the summed score for every fourth item (i.e. Parcel one: items 1, 5, 9, 13 and 17; Parcel two: items 2, 6, 10, 14 and 18 etc.). While item parcels can be useful in overcoming problems such as small sample sizes, one must be wary of their use when multidimensionality of the construct is a possibility (Little, Cunningham, Shahar, & Widaman, 2002). Given current uncertainty with regard to the dimensionality of the RPM tests, this could be problematic and may have influenced results. Hence, a similar study using individual items to form the factors, rather than item parcels, would be an interesting comparison to the results of Schweizer et al.'s study.

## Preliminary Conclusions: The RPM, Dimensionality and Gv

Taken as a whole, the evidence regarding the dimensionality of the RPM tests points to a unidimensional conceptualisation, although there exists sufficient evidence against this for questions regarding dimensionality to remain. Evidence from other research paradigms also indicates that it is worthwhile to continue to explore the role of Gv in RPM performance, despite the lack of support for this idea from the factor analytic and item response theory research. One potential explanation for the current status of conflicting results involves the common finding of sex differences in this test. If there are sex differences with regards to the latent structure of the APM, as suggested by Lim (1994), it would be problematic to analyse these two groups as one, and doing so would lead to spurious results. At this point, therefore, it is appropriate to include a discussion of reported sex differences in cognitive abilities and the implications that such differences may have for performance on the APM.

## Sex Differences in Cognitive Abilities

The existence or not of sex differences in cognitive abilities has been a controversial and well-researched topic throughout the history of psychology. Findings regarding this issue have been consistently influenced by sociopolitical goals, as well as personal beliefs, leading some to avoid the topic altogether. However, if the issues and biases regarding the research findings relating to sex differences in cognitive abilities are kept in mind, this research agenda can be useful in developing an understanding of how individuals solve certain problems and what kind of processes result in advantages or disadvantages on certain measures. Note that the term “sex”, rather than “gender”, is used here because it is the term that has traditionally been used in the study of cognitive abilities. However, the use of this term does not necessarily imply a biological cause for any apparent differences in cognitive abilities. This literature review also aims not to place any value judgment on differences in cognitive ability, but rather aims to be a discussion of the empirical evidence for the existence of such differences.

It is now generally agreed that no sex difference exists in *g* (Calvin, Fernandes, Smith, Visscher, & Deary, 2010; Camarata & Woodcock, 2006; Jensen, 1998; Mackintosh, 1996; but see Lynn, 1999 and Nyborg, 2005 for opposing arguments). Nonetheless, some consistent sex differences have been reported in certain broad abilities. For example, it is now widely accepted that males, on average, tend to perform better on tests of Gv, while females tend to show an advantage on speed of processing (Gs; Keith, Reynolds, Patel, & Ridley, 2008); the findings regarding Gf are less clear.

Although there is considerable debate in the literature regarding the causal factors of sex differences in abilities, most agree that these differences come about due to a complex interaction of psychological, social and biological factors (Halpern, 2012). While the concern of the present thesis is not to debate the developmental antecedents of sex differences in cognitive abilities, the existence of such differences in Gv and Gf will be explored, to the extent that these are relevant to the structural underpinnings of the RPM tests. Additionally, the notion of the existence of qualitative differences in thought processes (cognitive strategies), and the impact of these differences on test performance, will be considered.

## Methodological Issues in the Study of Sex Differences

Before embarking on a discussion of the types of abilities that have demonstrated sex differences, it is important to consider some of the general limitations of research in this area, which provide important caveats to many of the results reported. The majority of these limitations result from methodological issues in the statistical analysis, or sampling problems that result in conclusions being drawn from biased populations. Both of these issues will be considered in turn.

The particular statistical method used to investigate sex differences in ability measures has been shown to influence the outcome reported. There are several ways differences can be investigated; using composite scores (e.g. averages, weighted composites based on principal components analysis), using latent variables, or considering differences in the residual variable once a general intelligence factor has been partialled out. All of these methods can result in different conclusions (Keith et al., 2008). Keith et al. (2008) found varying patterns of sex differences depending on the type of analysis used: when a composite *g* score is used, a finding of a male advantage is more likely, whereas with the use of a latent *g*, slight female differences are found. However, the method used must also to some extent depend on what one wants to know. For example, differences in raw scores demonstrate that one group performs better on a particular test, which could be due to particular item features or test bias and are important for understanding raw test scores. Differences in the latent construct indicate that real differences may exist in the ability, assuming the inclusion of an appropriate variety of tests. These are different effects, and in interpretation, this must be taken into account.

Further, there is some evidence that findings of sex differences in broad abilities are influenced by sex differences in general intelligence; results differ when *g* is partialled out before sex differences are considered. Johnson and Bouchard (2007) found that when general intelligence variance was partialled out, sex differences in broad abilities became much stronger than previously found, while Lemos, Abad, Almeida, and Colom (2013) found that *g* largely accounted for the difference in numerical, abstract, verbal and spatial reasoning, but not mechanical reasoning. Whatever the case, it is apparent that the method of analysis chosen with regard to the partialling out or not of *g* variance can influence the results found.

Another issue with the study of sex differences is measurement invariance; without first establishing that the measure is invariant across groups, group differences can be meaningless if the intent is to examine sex differences in the ability, rather than test scores. Unfortunately, measurement invariance is not always established before group differences are investigated (Keith, Reynolds, Roberts, Winter, & Austin, 2011; Maitland, Intrieri, Schaie, & Willis, 2000).

Another methodological issue relating to the investigation of sex differences is sample selection. Often, studies use university samples to investigate sex differences, due to their easy access. However, such samples are often not representative of the population because the numbers of males and females recruited are unequal. Using a correction formula for the recruitment of individuals into such a sample, Hunt and Madhyastha (2008) demonstrated that conclusions can change substantially.

Another take on this problem is sample restriction in originally representative samples. Dykiert, Gale, and Deary (2009) reported findings indicating that participants who stayed in a longitudinal study which was initially representative of the population were more likely to be female or to have relatively higher IQs. As the authors argued, because males show more variance in their IQ scores, the combined effects of a higher ratio of female participants and higher overall IQ scores would result in a perceived male advantage in IQ.

Finally, there is substantial evidence of differences in the variability of both overall IQ scores and *g*-factor scores in males and females, with males more common at both the higher and lower ends of the distribution (Deary, Irwing, Der, & Bates, 2007; Johnson, Carothers, & Deary, 2008; Keith et al., 2008; Keith et al., 2011). This could result in a situation where, if only individuals of above average intellectual ability are selected into a study, the mean of male scores on a particular test may exceed that of females, simply because of their greater concentration at the higher end, when their greater concentration at the lower end is not also taken into account. This a feasible scenario when using samples from university populations. This variance difference has also been found in latent Gf (Keith et al., 2011; Lohman & Lakin, 2009; Strand, Deary, & Smith, 2006).

A final issue to keep in mind is the possibility that sex differences in both general and broad cognitive abilities may be developmentally dependent. As an explanation for some of the conflicting findings in the sex differences literature, Lynn (1999) proposed the developmental theory of sex differences in cognitive abilities, based on the notion that brain size is related to intelligence. According to this theory, girls mature in brain size at a faster rate than boys, resulting in a situation where girls' general intelligence overtakes that of boys at age 9, and girls continue to demonstrate an advantage on measures of general intelligence until around age 15. At around 15 to 16 years, the development of girls slows relative to boys and from this stage onwards boys score higher than girls on measures of general intelligence. Lynn (1999) and Colom and Lynn (2004) have presented results supporting this theory, but other research has not supported its predictions. For example, from an analysis of sex differences in a *g*-factor obtained from the Woodcock-Johnson III, Keith et al. (2008) reported developmental differences in *g*, but found that until age 18 there was no difference, and after age 18 there was a slight female advantage. Similarly, Reynolds, Keith, Ridley, and Patel (2008) found no significant age by sex interaction in a *g*-factor obtained from the Kaufman Assessment Battery for Children in a population aged 6 to 18 years. In light of findings outlined above about differences depending on whether scores or latent traits are examined, it is interesting to note that Lynn (1999) and Colom and Lynn's (2004) evidence comes from summed total scores on various intelligence test batteries, while Keith et al. (2008) and Reynolds et al. (2008) present latent variable results. While there may be developmentally related sex differences in cognitive abilities, it is unclear what exactly the pattern is. An alternative explanation for findings of no sex differences in childhood, but differences in adulthood, is that because children are much easier to access representatively, there tends to be less evidence for sex differences in abilities reported in children (Dykiert et al., 2009).

### Sex Differences in Gv

Given all the methodological issues discussed above, there remain a lot of unknowns in the area of sex differences in cognitive abilities. However, there are some very consistent findings, with little debate surrounding their existence; one such finding relates to sex differences in Gv. It is now a well-established finding that males, on average, perform better than females on certain tests of Gv. Large meta-analyses conducted over the last century have produced fairly consistent results supporting this contention (Linn & Petersen, 1985; Voyer et al., 1995). Although males do generally score higher on spatial ability tests, the effects are not homogeneous; certain types of Gv measures show a much larger male advantage than others, and some show no significant sex difference. The largest differences have consistently been found in tests of mental rotation (Linn & Petersen, 1985; Voyer et al., 1995). The effect size for sex differences in mental rotation is moderate to large, with a Cohen’s *d* between .56 (Voyer et al., 1995) and .73 (Linn & Petersen, 1985). Visualization measures tend to show the smallest sex difference, with results indicating that there is no significant difference between males and females (Linn & Petersen, 1985; Voyer et al., 1995). Linn and Petersen have argued that tests of visualization are distinguished from other tests due to the fact that they can be solved using multiple solution strategies, and that an analytic strategy is required for the most complex tasks. They have suggested that the distinction between Gv and Gf is particularly problematic for those tests classified as measuring visualization.

**Cognitive strategies.** There is little doubt that individuals use different strategies to solve problems. It is therefore possible that the use of different strategies when solving Gv test items may influence sex differences in Gv. If every strategy can be used equally effectively to solve all test items, then the issue of strategy is not problematic in terms of total scores. However, if certain test items are solved more readily using one strategy, then the test becomes not only a measure of the ability that it intends to test, but also of strategy. While different strategy use can influence performance on all types of tests, it has frequently been studied in relation to measures of Gv (see Kyllonen et al., 1984). Early research suggested that strategy may play a role in apparent sex differences in different types of Gv measures (Linn & Petersen, 1985). One of the most common distinctions drawn in the literature with regard to strategies and Gv is the distinction between holistic (or visual) and analytic (typically verbal) processes (Guay, McDaniel, & Angelo, 1978). Holistic processes refer to mainly visual strategies, which treat the stimulus as a whole, whereas analytic processes refer to mainly verbal-analytic processes, which treat the stimulus in a segmented way, often using verbal propositions. Gv tasks can be solved using different strategies, and there is some evidence to suggest that males tend to use an holistic strategy, while females tend to use a verbal-analytic strategy (Geiser, Lehmann, & Eid, 2006; Heil & Jansen-Osmann, 2008; Raabe, Höger, & Delius, 2006; Wang & Carr, 2014), which may contribute to the sex differences seen in these measures. Depending on the specific Gv task, one strategy may be more effective than another (Boulter & Kirby, 1994).

Although strategy differences are difficult to measure, methods such as examining the time taken to solve an item can provide clues as to which strategy is being used. For example, Raabe et al. (2006) found that on their specially designed rotation task, where an analytic strategy was expected to be more efficient, females performed the task more quickly than males, while Heil and Jansen-Osmann (2008) found that on their rotated polygon task, where an holistic strategy was expected to be more efficient, males performed the task more quickly, and the effect of stimulus complexity on solution time was weaker among males than among females. Other research suggests that while males may use a right-hemisphere, or spatial strategy, females show no hemispheric advantage, indicating they may capable of using two different strategies (Jaušovec & Jaušovec, 2012). In a study investigating strategy use on dynamic spatial tasks, Botella, Pena, Contreras, Shih, and Santacreu (2009) found that men tended to use an holistic strategy while women tended to use an analytic strategy, but men and women were represented in both strategy groups. Although there was still a significant sex difference in scores after accounting for strategy use, the effect size of this difference dropped substantially, indicating that differences in strategy use between males and females may be at least partially responsible for sex differences in spatial performance. Interestingly, among the group that used an holistic planned strategy, there was no sex difference in performance.

Not all studies, however, have shown a sex difference in strategy use (Miller, Donovan, Bennett, Aminoff, & Mayer, 2012) and some research suggests that sex differences in strategy use may only appear in participants of above average mental rotation ability (Jaušovec & Jaušovec, 2012), providing an explanation for the discrepant findings.

In attempting to understand how Gv and strategy differences might be related to RPM performance, the question of whether and what Gv ability is involved becomes important in understanding how this may affect both strategy use and sex differences. Although current research has provided some support for the notion that Gv is involved in RPM performance, it is not yet clear what the role of strategy differences may be, if any.

## Sex Differences in Gf

Findings regarding sex differences in tests of Gf have been inconsistent, reporting all possible outcomes (male advantage, female advantage and no difference). Regarding evidence using a latent Gf factor, findings tend to indicate no, or a negligible, sex difference (Arendasy & Sommer, 2012; Keith et al., 2008; Lakin & Gambrell, 2014).

However, studies that have considered sex differences in raw scores tend to report different results, depending on the type of test considered. A female advantage is often found on verbal Gf tests (Calvin et al., 2010; Lakin & Gambrell, 2014), although some evidence has suggested slightly superior male performance on verbal analogies tests (Colom, Quiroga, & Juan-Espinosa, 1999; Hyde & Linn, 1988; Lim, 1994). Although not strictly presented in a verbal format, letter series tests, which use alphanumeric stimuli (letters) to measure inductive reasoning, more commonly show a female advantage (Colom & Garcia-Lopez, 2002; Colom et al., 1999; Quereshi & Seitz, 1993; Rosen, 1995) or no difference (Codorniu-Raga & Vigil-Colet, 2003; Hakstian & Cattell, 1975b; Johnson & Bouchard, 2007; MacCann, 2010). Quantitative reasoning measures tend to show a male advantage (Calvin et al., 2010; Keith et al., 2008; Keith et al., 2011; Lakin & Gambrell, 2014). Number series, a combined measure of inductive and quantitative reasoning using numerical stimuli, has also shown a male advantage (Rosen, 1995; Steinmayr, Beauducel, & Spinath, 2010).

From these results, it is apparent that it is unclear whether there is a consistent sex difference in measures of Gf. The most consistent result is an absence of any sex differences in latent Gf, indicating no difference in the latent ability, but the presence of a male advantage in quantitative reasoning. It is possible that if a sex difference exists in measures of Gf, it may depend on the specific test. Beyond this, however, results are more complicated to interpret.

**Sex differences, Gf and figural stimuli.** One of the most popular formats of Gf tests is the figural format. This test format is popular for tests of inductive reasoning for the fact that inductive reasoning requires the individual to reason with novel stimuli, and figural stimuli tend to be seen as equally unlearned across individuals, while the use of verbal stimuli would involve prior knowledge to some extent, and such tests tend to demonstrate an additional loading on a Gc factor.

However, findings are again inconsistent with regard to sex differences in tests that use this format. The Cognitive Abilities Test figural battery has shown both a small, but significant female advantage (Calvin et al., 2010; Strand et al., 2006), a male advantage (Lakin & Gambrell, 2014) and no difference (Lohman & Lakin, 2009). The DAT-AR has shown a male advantage (Colom & Lynn, 2004; Colom et al., 1999), while the Culture Fair Intelligence Test (CFIT) has shown no difference (Colom & Garcia-Lopez, 2002) and the Differential Aptitude Scales (DAS) matrices subtest has demonstrated a female advantage (Keith et al., 2011). Interestingly, with regard to any developmental patterns, each possible result (male advantage, female advantage and no difference) was found in both child and adult samples, indicating that different developmental rates would not explain the difference. From these results, it is apparent that there is no consistent sex difference found across figural reasoning tests in general.

Steinmayr et al. (2010) considered sex differences in several tests of reasoning according to their content facet: verbal, numerical or figural. It was found that when sum scores were used, males showed an advantage on figural tests, but when latent scores on the figural factor were used this sex difference decreased substantially. These authors argued that this indicates that the male advantage on figural reasoning tests is due to a greater Gf capacity, rather than due to the particular type of stimulus, however this is inconsistent with findings indicating no sex difference in Gf.

Arendasy and Sommer (2012) found that despite the APM showing test bias in favour of males, there were no sex differences in the latent Gf factor, indicating that the test bias occurs due to some specific feature of the test or item design; conceivably, this could be a feature which introduces Gv into performance. One of the most interesting studies to consider sex differences in inductive reasoning specifically, used three different measures of inductive reasoning (CFIT, APM and PMA-R), and found different results for each test; no difference, male advantage and female advantage, respectively (Colom & Garcia-Lopez, 2002). This is inconsistent with Steinmayr et al.'s (2010) results, suggesting that, if anything, the figural content may be responsible for sex differences when tests of inductive reasoning specifically are used in analysis. Interestingly, the CFIT, which also utilizes figural stimuli, did show a small (*d* = .10) but non-significant male advantage, supporting this contention. It is unclear why it did not show such a large difference as did the APM, although it was originally argued by Cattell (1980) that, because the CFIT contains several subtests, specific variance could be cancelled out. It could be that there are specific Gv abilities involved in APM performance that are cancelled out in CFIT performance.

Colom and Garcia-Lopez's (2002) findings are potentially problematic for the suggestion that these three inductive reasoning tests measure the same underlying construct; if they are producing significantly different differences within the same sample, then this indicates that a substantial amount of variance is explained by a factor other than inductive reasoning. There is also other evidence to suggest that when sex differences in Gf test scores are aggregated across content domains, the differences are generally small or negligible, indicating that use of various content domains cancels out any substantial difference (Strand et al., 2006).

***Sex differences in the RPM.*** As with other tests of Gf, there are conflicting findings regarding sex differences in the RPM tests. There is a growing literature that has reported a male advantage on the APM, to the order of approximately .3 standard deviations (Abad et al., 2004; Colom, Escorial, & Rebollo, 2004; Colom & Garcia-Lopez, 2002; Lynn & Irwing, 2004a; Mackintosh & Bennett, 2005; Vigneau & Bors, 2008). However, many other studies have reported no difference, whether this be in mean scores (Arthur & Woehr, 1993; Bors & Stokes, 1998; Bors & Vigneau, 2001; Plaisted, Bell, & Mackintosh, 2011; Stephenson & Halpern, 2013) or according to IRT differential item functioning (DIF) methods (Chiesi, Ciancaleoni, Galli, Morsanyi, et al., 2012; Chiesi, Ciancaleoni, Galli, & Primi, 2012). Hence, it is unclear whether this difference does actually exist, and if it does, why it is only found in some studies.

Similar to findings regarding the APM, reports of sex differences are inconsistent with regard to the SPM. While several studies have found a raw score male advantage, either overall (Al-Shahomee, 2012), or on certain types of items (Lynn, Backhoff, & Contreras, 2005; Sellami, Infanzón, Lanzón, Díaz, & Lynn, 2010), others have found a female advantage (Abdel-Khalek, 1988), no difference (Flores-Mendoza et al., 2013; Pind, Gunnarsdottir, & Johannesson, 2003), or different results depending on participants’ ages (Bakhiet et al., 2015; Đapo & Kolenović-Đapo, 2012; Lynn et al., 2004). Interestingly, of all the above studies of sex differences in the SPM, none was conducted in a developed, western nation, the population for which these tests were designed, with results instead coming from Africa and Latin America. Administration of these tests in different cultures is potentially problematic (Owen, 1992). The one study that investigated sex differences in the SPM in the United Kingdom found no sex difference; Savage-McGlynn (2012) investigated both measurement invariance and differences in the latent mean, finding no evidence for any difference across the male and female groups.

Two recent meta-analyses regarding sex differences in the RPM, in both general and university populations, have found results supporting a male advantage (Irwing & Lynn, 2005; Lynn & Irwing, 2004b). However, an earlier extensive review by Court (1983) concluded that there was no sex difference in RPM scores. Hence, these large-scale studies have failed to settle the issue of the existence of sex differences in the RPM. Moreover, there have been criticisms of Irwing and Lynn's meta-analyses with respect to their selection of studies, the adequacy of their samples as representative, and choices in statistical procedures (Blinkhorn, 2005, 2006), casting some doubt on their conclusions. It is therefore clear from these conflicting results that more research is needed in order to determine whether there exists a sex difference in the RPM tests, either in raw scores, or at the latent mean level.

*The RPM, sex differences and Gv.* Given the questions regarding sex differences in the RPM, the hypothesis that Gv is involved in performance, and the well-established findings of male advantage in certain tests of Gv, it is important to consider if, and how, these points might be related. There is evidence from the literature regarding sex differences in the RPM that may help to understand the involvement of Gv processes in these tests. This research has not necessarily focused on differences in mean scores, but rather on differences in how males and females might go about solving items, and how performance on these tests might relate differentially to other variables, such as other cognitive ability test scores. Such sex differences may be able to tell us something about the involvement of Gv.

Firstly, with regard to the relationships between RPM performance and other measures, research by Geary, Saults, Liu, and Hoard (2000) suggests that there is a differential relationship between the sexes for visuospatial ability, as measured by the Mental Rotations test, and the RPM. These authors used structural equation modeling to examine whether the same pattern of structural relations would be found among several ability constructs, including Mental Rotation and the RPM, in males and females. For the female group, the standardized path coefficient from the RPM to Mental Rotation was much stronger (approximately double) than the coefficient between these two variables in the male group. However, Cockcroft and Israel (2011) found that the correlation between APM performance and verbal-analytic ability, as assessed by the Similarities subtest of the WAIS was not significantly different in males and females, suggesting that verbal-analytic ability contributes equally to APM performance across sex. Combined, these findings suggest that for females, Gv may be more important in RPM performance than for males, while verbal-analytic ability is equally important. This can be interpreted as congruent with Lim’s (1994) proposal that the APM measures two factors in females, whereas it measures only one in males.

A study using IRT DIF also found evidence for sex differences in the APM according to a visual/verbal-analytic distinction. IRT DIF is a method for detecting measurement bias by examining whether people with the same level of ability, but from different groups (e.g. sex), have a different probability of obtaining the correct answer on the same item. Abad et al. (2004) found that nearly 50% of items classified as requiring visual processes for solution were easier for males, while 50% of items classified as verbal-analytic were easier for females. However, only four verbal-analytic items were involved in this analysis, making interpretation of this second result problematic. There were, however, 11 visual items included in the analysis, and hence the results regarding these items are more robust. Additionally, these authors found that the sex difference in total APM scores decreased when those items displaying DIF were deleted, although a significant sex difference still remained. This result provides some support for the proposition that Gv is at least partially responsible for the average superior scores of males on the APM. Other evidence also points to Gv as a cause of the male advantage on APM. Colom et al. (2004) found that after controlling for scores on a test of spatial rotations, the sex difference in APM scores became non-significant. Finally, a study by Yang et al. (2014), using voxel-based morphometry, looked at patterns of brain activation while males and females where solving RPM problems (items taken from the SPM and CPM). They found different cortical activation patterns in men and women; males showed activation in the dorsolateral prefrontal cortex, which is associated with Gv, while females showed activation in the inferior frontal cortex, which is associated with verbal reasoning ability. Similarly, Njemanze (2005) found that, for correct answers, females showed left hemisphere activation and males showed right hemisphere activation, while for incorrect answers both sexes showed bilateral activation. The results of studies such as these suggest that there may be a differential relationship between Gv and the RPM between males and females, whether that is due to an actual difference in ability, or a difference in the strategies used to solve items.

**Cognitive strategies and the RPM tests.** Given the evidence for the use of different strategies in tests of Gv, it is plausible to consider whether such differences also exist in performance on the RPM tests. This is particularly true given the relatively commonly used distinction between visual and verbal-analytic RPM items, which have been associated with a visual-analytic strategy dimension. However, this issue has received less attention than has consideration of cognitive strategies utilized in performing Gv tasks. Brain imaging evidence described above, which found differences in brain activation according to sex, may relate to the use of different strategies in RPM performance; however, it is unclear whether it is strategy use, ability differences, or a combination of these factors that are responsible for these differences.

There have been a number of studies hypothesising a link between cognitive strategies and RPM tests (Hertzog & Carter, 1982; Lim, 1994); however, little research has investigated this link directly. Some evidence, however, does suggest that analytic strategy training can improve performance on the RPM tests. Kirby and Lawson (1983) found that analytic strategy training only resulted in gains on those items classified by task analysis as analytic items, while gestalt strategy appeared to be the default strategy, and gestalt items were equally likely to be solved correctly, whether students were trained in the gestalt or analytic strategy.

However, given the paucity of studies in the area, it is unclear exactly how strategy choice may be related to RPM performance, and how this might affect what the test measures. This strategy distinction can be particularly important in considering the question of what ability or abilities are measured by the RPM, because the use of different strategies can involve different abilities (Kyllonen et al., 1984). For example, Ozer (1987) found that, for both males and females, performance IQ was related to mental rotation performance, but that in females, verbal ability was also related.

## General Conclusions and Remaining Questions

Although there has been a substantial amount of research conducted with regard to both the structure of cognitive abilities and the psychometric properties of the RPM tests, it is still unclear what the relationship between Gf and Gv is, whether the RPM tests require the engagement of certain Gv abilities for successful performance, and what the role of sex differences in this relationship may be.

Although much evidence has indicated that the RPM tests are unidimensional, it is also clearly the case that findings have not been conclusive. There is some evidence for sex differences in the test, and it is therefore unclear whether the conclusions regarding unidimensionality and the utility of solution taxonomies may differ if males and females were considered separately. A study by Lim (1994) lends credence to this suggestion, finding that the APM loads onto only one factor in males, but two in females. If there were indeed a sex difference in the latent structure of the test, this could go some way to explaining some of the inconsistent findings in the literature to date.

Further, although considerable evidence established by factor analysis indicates that the RPM are likely unidimensional, as discussed, “unidimensionality” does not necessitate that only one ability is involved. Given that there is a substantial amount of research outside the literature concerning psychometric unidimensionality that suggests that Gv could be involved in RPM performance, this matter certainly requires further consideration. However, very few studies have considered directly the relationship of performance on measures of Gv to performance on the RPM, with the exception of Schweizer et al.’s (2007) study. However, given the limitations highlighted with that study, more research on this issue is needed.

Finally, given that there is some evidence that males outperform females on the RPM tests, together with some evidence that Gv abilities are involved in RPM performance, and considerable evidence that males perform better than females on certain tests of Gv, it is pertinent to investigate the interaction between sex differences and the role of Gv in RPM performance. Additionally, it is unclear if there exists a consistent sex difference in all tests of inductive reasoning (e.g. letter series, which tend to show a female advantage), or even in other figural matrices tests. Consideration of this question would allow us to investigate not only whether Gv might be responsible for any male advantage found in RPM performance, but also whether the type of test stimulus used is likely to influence the sex differences found. It may be the case that there is no consistent sex difference; however, a more thorough investigation of the existence of such differences could help to resolve this matter, and hopefully point to potential sources of this difference. On the basis of the literature, it is hypothesised that some characteristics of figural stimuli, as used in the RPM tests, are the source of a male advantage.

# Chapter 2: Exegesis

## Preamble

This body of research investigated the role of visuospatial ability (Gv) in performance on measures of fluid ability (Gf), particularly inductive reasoning, with a specific focus on the Raven's Progressive Matrices (RPM) series of tests. This was motivated by ongoing debate in the literature as to the dimensionality of the tests (Abad et al., 2004; Lynn et al., 2004; Vigneau & Bors, 2005) and the role of the figural stimuli used in these tests (Arendasy & Sommer, 2012). This was additionally motivated by consideration of visuospatial ability as a potential explanation for the often, although not always, reported male advantage on the RPM tests (Colom, Escorial, & Rebollo, 2004; Colom & Garcia-Lopez, 2002; Irwing & Lynn, 2005; Lynn & Irwing, 2004b). Therefore, a secondary and related aim was to examine the evidence for a sex difference in performance on measures of inductive reasoning, and to explore the implications of any such sex difference for both the practical and theoretical context of intelligence testing and theory.

Following is an outline of the key issues addressed in this thesis, including a discussion of how each of the included studies contribute to furthering understanding of the identified issues.

## Issue One: The RPM Tests

### Structure of the RPM Tests

Despite a large body of literature concerned with understanding the psychometric properties of the RPM tests, debate continues regarding the dimensionality of the tests, with various solutions proposed for both the Advanced version (unidimensional: DeShon et al., 1995; Abad et al, 2004; multidimensional: Dillon et al., 1981; Vigneau & Bors, 2005) and the Standard version (unidimensional: van der Elst et al., 2013; multidimensional: Grigoriev & Lynn, 2014; Lynn et al., 2004; van der Ven & Ellis, 2000). Furthermore, substantial evidence from other research paradigms indicated the involvement of Gv in addition to Gf in the RPM tests, leading to additional questions surrounding the dimensionality of the tests.

Some of the variation in factor analytic results could be explained by the use of inappropriate methods of data analysis for the type of data obtained from the RPM tests. However, another potential cause of the disparate results regarding the factor structure of the RPM considered in the present case was whether the factor structure was the same across sex. Research by Lim (1994) indicated some differences in the factor structure of intelligence between males and females, and particularly differences in the abilities required for successful performance on spatial analogies tasks, although other studies have reported little sex difference in the factor structure of intelligence (Hertzog & Carter, 1982; Reynolds et al., 2008; Rosen, 1995). For some time there has also been a suggestion that different strategies can be used to solve the RPM items (e.g. Hunt, 1974), and there is evidence to suggest that males and females may use different cognitive processes or strategies in solving APM problems and visuospatial test items more generally (Geiser et al., 2006; Heil & Jansen-Osmann, 2008; Mackintosh &

Bennett, 2005; Plaisted et al., 2011). If this were the case, these factors could contribute to a lack of measurement invariance across sex.

Measurement invariance had been confirmed in the SPM Plus for children and adolescents (Savage-McGlynn, 2012), but not in the APM. Chiesi, Ciancaleoni, et al. (2012) tested unidimensionality of Arthur and Day's (1994) short form APM in separate male and female samples, but did not directly test measurement invariance. If the APM did indeed violate measurement invariance across sex, this could not only explain contradictory results concerning the factor structure of the test, but would have serious implications concerning making comparisons across sex, and in understanding what ability or abilities the test is measuring. If measurement invariance is not met, then it makes little sense to draw conclusions about the test in a combined group of males and females. This motivated the establishment of measurement invariance across sex before further investigation could continue.

Furthermore, while previous studies had investigated the suitability of solution taxonomies proposed by DeShon et al. (1995) and Carpenter et al. (1990) for explaining the latent structure of the APM (Abad et al., 2004; Vigneau & Bors, 2008), these studies had only considered sex differences at the item level, using IRT DIF procedures and composite scores, respectively. Therefore, despite previous research indicating a one-factor model was a better fit to the data than factor solutions based on the solution taxonomies, to our knowledge, this comparison had not been investigated in males and females separately.

If measurement invariance was confirmed, we were also interested in investigating the presence of any latent mean differences in the construct measured by the APM. Although some previous research has reported sex differences in manifest scores, it was unclear whether this difference would also be found at the latent level.

Latent mean differences tell us something different to what differences in manifest scores tell us; a difference in manifest scores may be due to specific characteristics of certain items, that is, an artifact of measurement, while differences in latent means take into account measurement error and determine whether there is a difference in the underlying construct measured (see e.g., Keith et al., 2008).

**Methodological Issues.** As previously stated, less than ideal methods of data analysis have frequently been used in factor analyses of the RPM tests. This is partly due to the historical difficulty of performing factor analyses with binary data, because of poor computing power and a lack of advances in statistical understanding.

However, more recently, newer and more appropriate methods of analysis have become available and more accessible. Analysis of binary data should ideally be performed with tetrachoric correlations; the use of Pearson’s correlations can attenuate the correlation because of violations of the assumptions of continuous variables and a bivariate normal joint distribution. The weighted least squares mean and variance adjusted estimator (WLSMV) available in Mplus (L. K. Muthén & Muthén, 1998-2012) uses tetrachoric correlations and has been shown to perform well with binary data (B. Muthén, du Toit, & Spisic, 1997). However, this method has only been used to examine the factor structure of Arthur and Day’s (1994) short form of the APM (Chiesi, Ciancaleoni, Galli, Morsanyi, & Primi, 2012), but not the full form, nor other short forms. It should be noted, however, that other methods such as LISREL’s DWLS have been used to examine the structure of the full APM (Abad et al., 2004). DWLS is similar to WLSMV; however, the two estimation methods use different asymptotic approximations to estimate the covariance matrix and produce slightly different results.

Additionally, and pertinent to the present case, the WLSMV procedure may be advantageous with multiple groups models (B. Muthén & Asparouhov, 2002).

Furthermore, although Item Response Theory (IRT) has been around for many years now, Classical Test Theory (CTT) has typically dominated within the intelligence literature, and in examining the dimensionality of the RPM tests. IRT was designed for analysis of binary data, and is therefore ideal for that purpose. IRT has several advantages over CTT, including that statistics are not sample dependent and that it is considered item oriented, rather than test oriented. More specifically, it allows predictions to be made regarding an individual's probability of answering an item correctly given the item characteristics. While CTT considers the test score as a combination of the effects of the latent trait and error, IRT considers the probability of obtaining a correct solution to a particular item given individual ability and item difficulty, and in some cases item discrimination and guessing; therefore, these two approaches conceptualise measurement somewhat differently. However, IRT is not appropriate in all situations, given its stronger assumptions regarding the characteristics of the data and the requirement for large sample sizes.

Of the IRT models, the Rasch model has the most desirable measurement properties (Embretson & Reise, 2000). It is also the most well established, well understood and well researched. This model considers the probability of a correct response to a particular item as a function of both the individual's level of ability and the item difficulty. Rasch model estimates tend to be more stable than other IRT models given that only the single parameter of item difficulty requires estimation, unlike the 2PL and 3PL models which require estimation of two and three parameters, respectively. Additionally, the Rasch model is the only IRT model that has the scaling properties of linear, interval measurement (Embretson & Reise, 2000).

One of the assumptions of the Rasch model is unidimensionality of the data, and therefore the fit of this model to the data can be used to determine how well the data conform to this assumption. A further argument for the Rasch model is that it is less clear how the 2PL and 3PL models relate to the issue of unidimensionality. The 2PL and 3PL models consider additional parameters besides difficulty: discrimination in the case of the 2PL model and discrimination and guessing in the case of the 3PL model. Although this may be useful in understanding how items perform, some proponents of the Rasch model argue that these models are overparameterized and they may also contain problematic measurement features. For example, it is suggested that if discrimination is allowed to vary across items then the meaning of the construct being measured changes along with these changes in discrimination, which is problematic for true measurement of a specific construct (Salzberger, 2002). As maintained by Vigneau and Bors (2005), the addition of other parameters besides difficulty introduces an additional source of variance into item performance beyond ability, which may result in items differing in their correlation with the total score, something that occurs because they are not measuring the same thing.

Differential item functioning (DIF) under IRT modeling can be used to investigate variations in the performance of different groups with the same level of ability on the same item. Some work had been conducted on the APM within the framework of IRT, however, only one study had used the Rasch model (Vigneau & Bors, 2005), and this study had not examined DIF.

Although CFA using the WLSMV estimator and Rasch analysis are mathematically quite similar, their theoretical background and approach to measurement differ. Therefore, they represent two distinct theoretical approaches for examining the structure, dimensionality, invariance and differential item functioning of a measure and can be used in a complementary way. The use of both methods in investigating the dimensionality and measurement invariance of the APM was seen as advantageous in the present case.

### Validity of the RPM Tests

Despite claims that the RPM tests are among the best single measure of *g*, there continue to be arguments for the involvement of visuospatial ability in performance on these tests (Colom, Escorial, & Rebollo, 2004; DeShon et al. 1995; Ebisch et al., 2012), although some present evidence against this (Schweizer et al., 2007). This notion is a threat to the construct validity of the tests; if these tests do indeed involve Gv to a substantial extent, then they must consequently not be particularly good measures of Gf, as we understand this construct. Given the conflicting reports in the literature, this issue deserved further attention.

### New Contributions

**Paper 1.** In order to address the issues regarding the structure of the RPM, Paper 1 examined the dimensionality of three different versions of the APM among three different samples using both CTT (CFA using the WLSMV estimator) and IRT (Rasch modeling). Comparison of the results obtained from CTT and IRT methods allowed for a stronger conclusion regarding the dimensionality of the APM.

Differences in the factor structure and latent means were investigated across sex using Multiple Groups Confirmatory Factor Analysis (MGCFA). Measurement invariance was important to establish before further investigation; if the APM were found not to be measurement invariant across sex then it would make little sense to continue analysing and comparing combined samples of males and females.

**Paper 2.** In order to address the validity of the RPM tests as a measure of Gf, structural equation modeling was used to examine the incremental validity of Gv in predicting performance on the RPM over alternate measures of Gf. This paper extended the work of Schweizer et al. (2007).

## Issue Two: Intelligence Theory

### The relationship between Gf and Gv

The CHC model of intelligence (see Schneider & McGrew, 2012, for a recent review) proposes a taxonomy of cognitive abilities, including Gf and Gv at the broad ability level. Although the definitions of these abilities are conceptually distinct, in practice the distinction is much less clear, particularly with regard to the narrow Gf ability Induction. In order to measure Gf, the use of stimuli which are either novel or over-learnt is required and this has traditionally led to the use of abstract figures in common measures of Gf, and particularly measures of Induction (the other two Gf factors of sequential reasoning [RG] and quantitative reasoning [RQ] are more likely measured with verbal and numerical stimuli, respectively). It is somewhat unclear what the impact of this is; it is not uncommon for factor analytic research to identify a combined Gf/Gv factor (Crawford, 1991; Salthouse et al., 2008), and Carroll’s (1993) analysis found substantial overlap between certain Gf and Gv measures. J. L. Horn (1988) argued that the distinction between Gf and Gv is in the degree of reasoning demands in the task as compared to the degree of visualization demand. While this argument gives a good conceptual idea of what a Gv measure may involve compared to a Gf measure, it still leaves unclear the extent to which the two abilities are involved in certain measures, while suggesting that substantial overlap will be present among measures of Gf utilizing visual stimuli.

Therefore, an understanding of the extent of involvement of Gv abilities in performance on measures of Induction is important not only in furthering understanding of what abilities particular tests measure, but also in furthering understanding of the relationship between Gf and Gv.

**The role of the content facet.** Related to the issue of the relationship between Gf and Gv is consideration of the role of the content facet in the ability measured. The BIS model highlights the importance of the “content” of the test in addition to the “process” by classifying every component of intellectual ability as depending on these two elements. This is certainly relevant to a discussion of the relationship between Gf and Gv. Broadly speaking, Gf can be likened to processing capacity, a cognitive “process” (also known as reasoning), while Gv may be likened to figural content. Under the BIS, the RPM tests are classified as figural reasoning (Süβ & Beauducel, 2015). Although the process facet has been found to account for more variance than the content facet, the content does account for non-trivial variance (Bucik & Neubauer, 1996). This model is particularly appealing when considering issues related to the role of Gv in RPM performance, and issues of the relationship between Gv and Gf broad abilities; it provides a framework for understanding this relationship a little better. In fact, Wilhelm (2005) suggests that the distinction between RG and inductive reasoning is actually a distinction between the content used.

### New Contributions

**Paper 2.** Paper 2 addresses the issue of the relationship between Gf and Gv within the specific context of the RPM tests. It allowed examination of whether Gv abilities did in fact contribute to performance on the RPM, a widely accepted measure of Gf. A large part of the overlap between Gf and Gv abilities could be due to the similar format of Gv measures, and many inductive reasoning measures. By including a measure of Gf that did not utilize figural stimuli in order to account for the Gf variance in the RPM, the unique contribution of the Gv abilities to RPM performance over and above Gf could be examined.

**Paper 3.** Paper 3 addresses this issue in its consideration of sex differences in different tests of inductive reasoning, and whether the content used in the test can account for some, or all, of the sex differences identified. Although this study was not directly able to examine the relationship between Gf and Gv, it was hypothesized that the involvement of Gv in many measures of inductive reasoning (i.e. those utilizing figural stimuli) might cause a male advantage. By comparing measures of inductive reasoning that utilized different stimuli, we could uncover patterns in the sex differences observed.

**Paper 4.** During searches for data pertinent to sex differences in measures of inductive reasoning for Paper 3, a large set of Australian data for the General Reasoning Test 2 was obtained. This test consists of three subtests: abstract reasoning, verbal reasoning and numerical reasoning. The abstract reasoning subtest is similar to the RPM tests in its use of figural stimuli, but uses a different format; rather than matrices, it uses a combination of series, classification and analogy items. Little published work had previously been conducted on this test, and no work had been conducted within the Australian population, despite its relatively widespread use in applied organisational settings.

The item format of this test brought up the possibility of investigating not only the structure of the reasoning test, but also what impact the different item stimuli might have on the measurement characteristics of the test and on its latent structure.

Paper 4 addresses the structure of this reasoning test and the role of stimulus type by considering the role of both the “content” and the “process” of each item. This paper also considers how abstract (figural) reasoning items are related to verbal and numerical reasoning items in examining the factor structure of the test.

## Issue Three: Sex Differences

Another long-standing issue within the intelligence field has been sex differences, both in general intelligence, and in specific abilities. Although general consensus suggests there is no difference in general intelligence (Calvin et al., 2010; Camarata & Woodcock, 2006; Keith et al., 2011; Mackintosh, 1996), some researchers propose a male advantage (Lynn, 1994), and often use scores on the RPM tests to support this claim (Irwing & Lynn, 2005; Lynn & Irwing, 2004b), although not always (Colom & Lynn, 2004; Irwing, 2015). It thus becomes important to consider why sex differences on the RPM tests occur.

Although research findings do suggest sex differences in certain specific abilities, it is unclear whether such a difference exists in measures of Gf. Although Lynn and Irwing (2004b) and Irwing and Lynn (2005) report a male advantage on the

RPM tests, other measures have not necessarily shown the same result (Calvin et al., 2010; Colom & Garcia, 2002, Keith et al., 2011).

Furthermore, given the proposed role of Gv in performance on the RPM tests, and the established male advantage on Gv measures, it is prudent to examine whether the sex difference reported in the RPM tests is found more broadly in Gf tests in general. It is also prudent to examine whether Gv may account for the sex difference in measures of Gf that demonstrate a male advantage. Research by Colom and Garcia-Lopez (2002) has found different effect sizes of the sex difference depending on the measure of inductive reasoning used, while research by Colom, Escorial and Rebollo (2004) suggests that Gv can account for the observed difference. However, Colom and Garcia's results were based on a single sample, while Colom, Escorial and Rebollo considered this issue at the manifest score level.

## New Contributions

**Paper 1.** Paper 1 examined evidence for measurement invariance and latent mean differences across males and females in several forms of the APM, thereby expanding on results concerning sex differences in manifest scores on this measure.

**Paper 2.** In order to further address the issue of sex invariance considered in Paper 1, Paper 2 examined structural invariance of the relationships between Gf, Gv and RPM across sex. This involved examination of whether the relative contribution of Gf and Gv abilities to performance on the Advanced RPM was invariant, motivated by the notion that males and females may use different strategies to solve RPM items, and therefore different abilities may be involved to varying extents. This allowed for investigations beyond whether the latent APM factor was the same across sex, by examining whether the same abilities contributed to the same extent. Measurement invariance as studied in Paper 1 allowed us to determine that the construct measured by the APM was invariant, but was unable to provide information regarding what that construct might be.

Paper 2 also addresses the issue of a sex difference in the RPM tests and the role of Gv by considering whether the sex difference in Gv accounts for the sex difference found in the APM. This was done at the latent level, rather than at the manifest score level.

**Paper 3.** Paper 3 addresses the issue of sex differences in inductive reasoning by examining, firstly, whether sex differences do indeed exist in manifest scores on measures of inductive reasoning, and secondly, whether the existence, magnitude and direction of the difference varies as a function of the specific type of test considered. Again, this allowed consideration of how the type of stimulus used might influence the performance of these two groups and, consequently, could be an influencing factor in what the test measures. By using meta-analytic techniques to summarise a large amount of data, conclusions drawn from these findings are stronger than conclusions drawn from a single sample.

**Paper 4.** This paper addressed the issue of sex differences in the General Reasoning Test 2. Measurement invariance across sex was investigated, as were sex differences in the latent mean and at the individual item level.

## Issues Beyond the Scope of this Thesis

It is acknowledged that there is growing interest in the relationship between working memory and Gf (Ackerman, Beier, & Boyle, 2005; Chuderski, 2015; Colom et al., 2015; Harrison, Shipstead, & Engle, 2015; Martínez et al., 2011; Zeller, Wang, Reiß, & Schweizer, 2017) and that there is evidence that goal management is a key factor in RPM performance, likely related to working memory (e.g. Loesche, Wiley & Hasselhorn, 2015; Primi, 2014). However, this thesis was specifically focused on the relationship between Gv and Gf in the context of the RPM and other inductive reasoning measures. As such, an in-depth consideration of working memory was beyond the scope of this thesis.

Furthermore, this thesis is largely focused on research within the psychometric tradition. Although some consideration of cognitive processing accounts of reasoning is given (e.g. Carpenter et al., 1990; Kunda, McGreggor, & Goel, 2013; Primi, 2001), these are primarily considered in the context of the relationship between Gf and Gv.

# Chapter 3: Paper 1

## Preamble

The rationale for Paper 1 came from three main sources: Firstly, previous research on taxonomies of solution rules required to solve RPM items and how these solution taxonomies may relate to the factor structure of the RPM tests; secondly, methodological issues and conflicting results in previous factor analyses of the RPM tests; and thirdly, a lack of research into sex differences at the latent level.

With regard to the first point, various solution taxonomies have been proposed for solving the items of the RPM tests, with a main focus on the Advanced version. These solution taxonomies have considered the types of rules governing the relationships between elements of the stimuli in each figural matrix. Rules are repeated across different items, and have been classified as visuospatial or analytic by some authors (DeShon et al. 1995), or analytic only by others (Carpenter et al. 1990).

The first of the popular solution taxonomies was developed by Carpenter et al. (1990). This analysis resulted in the creation of five rules that can be used to solve 34 of the 36 items in the APM. These rules were presented in Table 1.2. Following Carpenter et al.’s (1990) analysis, DeShon et al. (1995) sought to differentiate between rules or items involving a visual strategy and those involving an analytic strategy. As Carpenter et al.’s (1990) rules all represented an analytic strategy for solving items, DeShon et al. modified this taxonomy to include additional visual strategies. Table 1.3 presented the rules provided by DeShon et al. As can be seen in Table 1.3, the analytic rules of DeShon et al. (1995) tend to correspond with the rules of Carpenter et al. (1990), with the exception of the addition/subtraction rule which was placed under the heading of visual strategies.

Both of these taxonomies have proved popular in analysing and understanding the processes involved in solving the APM items; however, there is some confusion regarding how these two models might fit together. Given the disparate literature on the classification of solution rules, a preliminary step before conducting the analysis in Paper 1 was to consider whether the existing rule taxonomies could be synthesised. Following is a description of this preliminary work concerning the potential to condense and synthesise solution taxonomies.

**Rule Synthesis**

Firstly, because DeShon et al.'s (1995) taxonomy contains 10 different rules, significantly more than Carpenter et al.'s (1990) 5 rules, the rules of DeShon et al. (1995) were inspected to determine whether any of them appeared to represent a similar process. The first finding was that three of DeShon et al.'s (1995) visual rules could be combined into one: superimposition and superimposition with cancellation could be considered special types of the addition/subtraction rule. Figure 3.1 demonstrates the equivalency of these rules.

Figure 3.1(a) shows an example of the addition/subtraction rule, where the elements of the first square are added to the elements of the second square to obtain the third square. The example shown in Figure 3.1(b) demonstrates the superimposition rule. However, an addition/subtraction rule could equally apply, in that the features of the first and second squares are added together to obtain the third square. Figure 3.1(c) provides an example of the superimposition with cancellation rule. As with Figure 3.1(b), an addition/subtraction rule could also apply. For example, to form the third square, the elements of the first and second squares are added together, and then those elements that the first two squares have in common are subtracted.

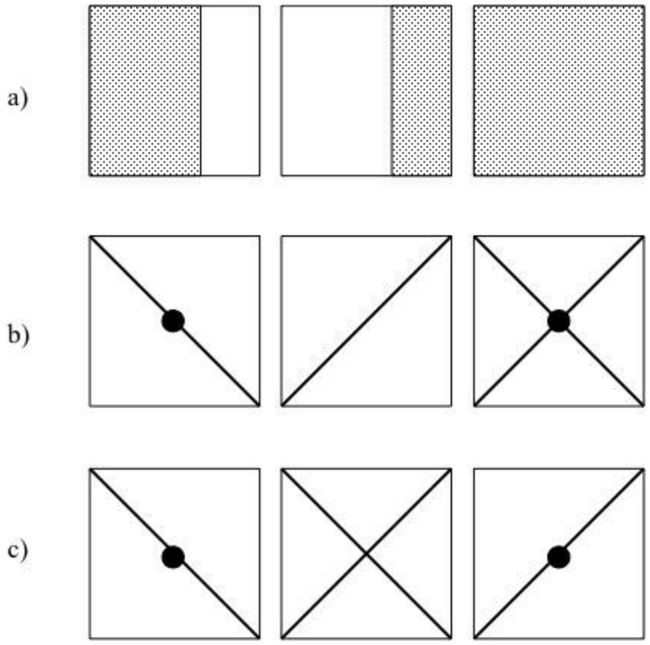


*Figure 3.1.* a) Addition/Subtraction; b) Superimposition; c) Superimposition with cancellation.

Another rule from DeShon et al. (1995) that may be considered a form of the addition/subtraction rule is the rule of superimposition with conditional placement. This rule occurs only three times in the APM, and may be a more analytic form of addition/subtraction, given its somewhat more complex rule system. Figure 3.2 shows an example of this rule. The entire matrix is given, as an understanding of all rows is necessary to obtain the correct answer for the bottom right hand square.

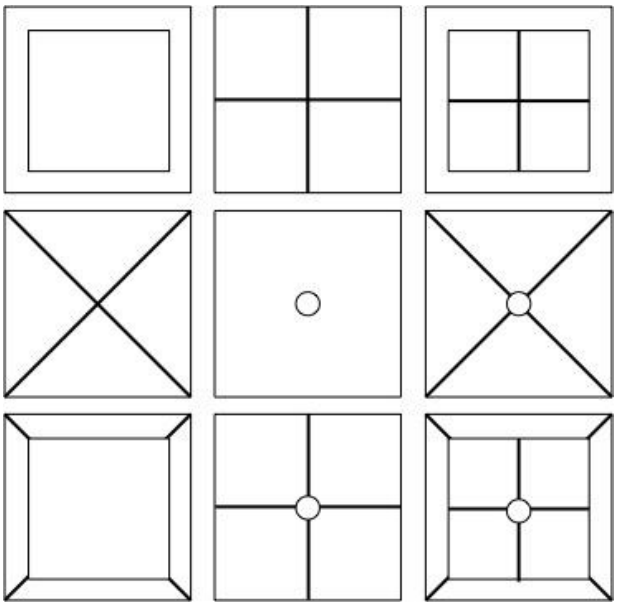


*Figure 3.2.* Superimposition with conditional placement.

As can be seen in Figure 3.2, each row and column involves addition of the elements in the first two squares to obtain the third square. However, additionally, the conditional placement must be noted, in that for the cross, the lines appear only within the small square, while for the X, the lines appear only outside the small square.

Further inspection of DeShon et al.'s (1995) rules indicated that two of their visual rules could be combined with one of their analytic rules: movement and rotation could be thought of as a special case of the quantitative pairwise progression rule. Figure 3.3 demonstrates the equivalency of these rules.

Figure 3.3(a) shows an example of the quantitative pairwise progression rule, where there is a quantitative decrement in the length of the horizontal block across the row. Although Figure 3.3(b) represents an example of the movement rule, the quantitative pairwise progression rule could equally apply in that there is a quantitative increment in the position of the patterned circle across the row. Similarly, although Figure 3.3(c) represents an example of the rotation rule, the quantitative pairwise progression rule could equally apply in that there is a quantitative change in position of the triangle (rotated 90 degrees) across the row.

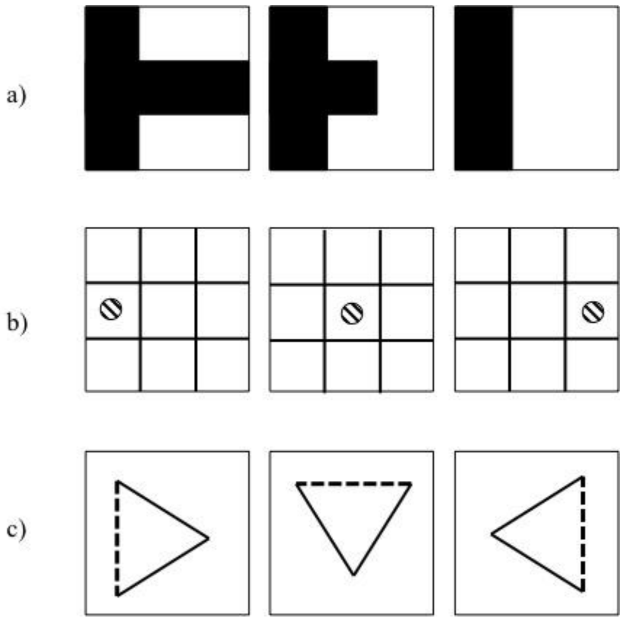


*Figure 3.3.* (a) Quantitative pairwise progression; (b) Movement; (c) Rotation.

**Results of Rule Synthesis**

Once these changes are applied to DeShon et al.'s (1995) taxonomy, the near equivalence of the two rule classification systems can be observed (Table 3.1). Out of the 32 items that are classified under both taxonomies, 81% of the item rules are now in agreement. By inspecting some of those items still in disagreement, this percentage can be improved. For example, items 22 and 23 are classified by Carpenter et al. (1990) as distribution of two and by DeShon et al. (1995) as addition/subtraction. These items were originally classified as superimposition with cancellation by DeShon et al. (1995). All other items originally classified as superimposition with cancellation are represented by the addition/subtraction rule of Carpenter et al. (1990), and not the distribution of two rule (items 9, 12, 16 and 33). By inspecting the characteristics of items 22 and 23 and comparing these to items 9, 12, 16 and 33 on one hand, and items 30-32 and 35-36 on the other, it is apparent that the rule necessary to solve these two items corresponds more closely to the former. Hence, it will be argued that items 22 and 23 represent the addition/subtraction rule. This results in an 88% agreement across the two taxonomies.

In terms of the visual-analytic split, it can be seen that all distribution of three and distribution of two items are analytic. Eight (73%) of the addition/subtraction items are visual (23% involve both visual and analytic processes). The picture is more complicated with regard to quantitative pairwise progression. If we exclude those items where quantitative pairwise progression operates in conjunction with an analytic rule (distribution of two/distribution of three), four items are visual, one is analytic, one is both, and five are either. This indicates that the quantitative pairwise progression may be able to be performed in either a visual or an analytic manner. Furthermore, the visual versus analytic classifications with regard to quantitative pairwise progression do not appear to be the result of the items originally classified as movement or rotation representing a visual factor and those originally classified as quantitative pairwise progression representing an analytic factor, as the only two items originally classified as quantitative pairwise progression were not only analytic.

Table 3.1

*Comparison of Modified DeShon et al. (1995) with Carpenter et al. (1990)*

| Item | Carpenter et al. (1990) and Mackintosh & Bennett (2005) | Modified DeShon et al. (1995) | Verbal/Analytic Classification (DeShon et al., 1995) | Original DeShon et al. (1995) |
|---|---|---|---|---|
| 1 | D3 | D3 | Analytic | D3 |
| 2 | - | P | Either | E or P |
| 3 | P | P | Visual | M |
| 4 | P | P | Analytic | P |
| 5 | P | P | Either | P or Constant |
| 6 | P | P | Either | M or P |
| 7 | A/S | A/S | Visual | SI |
| 8 | D3 | D3 | Analytic | D3 |
| 9 | A/S | A/S | Visual | SIC |
| 10 | P | P | Visual | M |
| 11 | - | A/S | Visual | A/S |
| 12 | A/S | A/S | Visual | SIC |
| 13 | D3 | D3 | Analytic | D3 |
| 14 | P | P | Either | M or P |
| 15 | A/S | - | - | - |
| 16 | A/S | A/S | Visual | SIC |
| 17 | D3 | D3 | Analytic | D3 |
| 18 | - | MT | Visual | MT |
| 19 | A/S | A/S | Both | SI with CP |
| 20 | A/S | A/S | Both | SI with CP |
| 21 | D3 | D3 | Analytic | D3 |
| *22* | *D2* | *A/S* | *Visual* | *SIC* |
| *23* | *D2* | *A/S* | *Visual* | *SIC* |
| 24 | P | P | Visual | M |
| 25 | P | P | Both | M with CP |
| 26 | P, D3 | P, D3 | Both | R with D3 |
| 27 | D3 | D3 | Analytic | D3 |
| 28 | D3 | D3 | Analytic | D3 |
| 29 | D3 | D3 | Analytic | D3 |
| 30 | D2 | D2 | Analytic | D2 |
| *31* | *D3, D2* | *P, D3, D2* | *Both* | *M & D3, D2* |
| *32* | *D3, D2* | *P* | *Visual* | *M* |
| *33* | *A/S* | *P, A/S* | *Visual* | *R & SIC* |
| 34 | D3 | D3 | Analytic | D3 |
| *35* | *D2* | *A/S* | *Both* | *SI with CP* |
| 36 | D2 | D2 | Analytic | D2 |

*Note*: P = quantitative pairwise progression, A/S = addition/subtraction, D3 = distribution of three, D2 = distribution of two, M = movement, R = rotation, SI = superimposition, SIC = superimposition with cancellation, SI with CP = superimposition with conditional placement, E = expansion

As noted by Carpenter et al. (1990), it may also be possible to combine some of their rules, for example, quantitative pairwise progression may be reducible to distribution of three, and addition/subtraction may be sufficient to solve distribution of two items. Mackintosh and Bennett (2005) argue, therefore, that quantitative pairwise progression and distribution of three map onto DeShon et al.'s (1995) analytic items, while addition/subtraction and distribution of two map onto DeShon et al.'s visual items. However, as this work has shown, several of the quantitative pairwise progression items may involve visual processes. Furthermore, when items 22 and 23 are classified as addition/subtraction, the distribution of two items no longer represent DeShon et al.'s visual items, which would indicate that distribution of two may measure something quite different to addition/subtraction.

By synthesising the work of both Carpenter et al. (1990) and DeShon et al. (1995), we can minimise the number of rules involved in APM solutions, thereby obtaining a simpler taxonomy. Furthermore, although the taxonomy then becomes very similar to the original taxonomy provided by Carpenter et al. (1990), the visual-analytic distinction provided by DeShon et al. (1995) provides additional information.

### Dillon et al.'s (1981) Factor Analysis

Although not a taxonomy of solution strategies, Dillon et al.'s (1981) factor analysis of the APM has been influential. Dillon et al. identified two factors in the APM: pattern progression and addition/subtraction. Pattern progression was defined as the ability to perceive a recurring or sequential design. Hence, this factor may represent the quantitative pairwise progression and distribution of three rules. Quantitative pairwise progression can be considered a pattern in that the quantitative increment or decrement is a type of sequence. Similarly, distribution of three can be considered a pattern in that the three features are often displayed in a certain sequence or recur across rows in the APM items. Addition/subtraction was defined as represented by items where solution requires addition or subtraction of elements. Hence, this factor may represent the addition/subtraction rule. Table 3.2 presents the items identified by Dillon et al. as good representations of the pattern progression and addition/subtraction factors along with these items' corresponding rule taxonomy classification.

Table 3.2

*Dillon et al.'s (1981) item types*

| Item | Dillon et al. | Rule |
|---|---|---|
| 2 | PP | P |
| 3 | PP | P |
| 4 | PP | P |
| 5 | PP | P |
| 7 | A/S | A/S |
| 9 | A/S | A/S |
| *10* | *A/S* | *P* |
| 11 | A/S | A/S |
| 16 | A/S | A/S |
| 17 | PP | D3 |
| *21* | *A/S* | *D3* |
| 26 | PP | P, D3 |
| *28* | *A/S* | *D3* |
| 35 | A/S | A/S |
| *36* | *PP* | *D2* |

*Note*: P = quantitative pairwise progression, A/S = addition/subtraction, D3 = distribution of three, D2 = distribution of two, PP = pattern progression.

As can be seen in Table 3.2, Dillon et al.’s (1981) classification does broadly correspond to a distinction between quantitative pairwise progression and distribution of three on the one hand, and addition/subtraction on the other hand. Six of the seven items (86%) classified as pattern progression correspond to either the quantitative pairwise progression or distribution of three rule. However, only five of the eight items (63%) classified as addition/subtraction correspond to the addition/subtraction rule. When the other three items are inspected, it is difficult to determine how their solution may involve any type of addition or subtraction. Hence, Dillon et al.’s interpretation of this factor may be questioned.

The combination of the quantitative pairwise progression and distribution of three rules is also difficult to interpret given their clear separation in Carpenter et al. (1990) and DeShon et al.’s (1995) taxonomies. However, given the interpretation of distribution of three as analytic, and the interpretation that quantitative pairwise progression may be either analytic or visual, it may be that the pattern progression factor represents those items that involve an analytic form of quantitative pairwise progression. Consistent with this interpretation, only one of the items identified by Dillon et al. (1981) as pattern progression and by Carpenter et al. and DeShon et al. as quantitative pairwise progression was not classified as involving analytic, either, or both processes by DeShon et al. Similarly, the quantitative pairwise progression item classified by Dillon et al. as addition/subtraction was classified by DeShon et al. as visual.

Therefore, although Dillon et al.’s (1981) factors do not match perfectly the solution rules of Carpenter et al. (1990) and DeShon et al. (1995), there is some evidence to suggest it may correspond to DeShon et al.’s visual-analytic distinction.

## Preliminary Conclusions from Solution Taxonomy Synthesis

Results of the synthesis of solution taxonomies indicated that the number of solution rules proposed by DeShon et al. (1995) could be condensed, and these condensed rules were used in the subsequent factor analysis performed in Paper 1. Results also indicated that DeShon et al.'s and Carpenter et al.'s (1990) solution taxonomies were far more similar than they initially appeared, with the classifications reaching 88% agreement. The main distinction between these two taxonomies then becomes that one highlights the visual-analytic distinction, while the other simply describes the different types of solution rules. Additionally, there was some evidence that Dillon et al.'s (1981) factors may broadly correspond to DeShon et al.'s visual-analytic distinction, although there were some differences. Therefore, included in Paper 1 were models based on DeShon et al.'s visual-analytic distinction, Carpenter et al.'s rules and Dillon et al.'s factors.

Although previous work has failed to find support for these taxonomies as representative of different latent factors in the APM (e.g. Abad et al., 2004; Vigneau & Bors, 2008), research has not examined the factor structure separately by sex and, furthermore, other research does indicate that the rules contained in these taxonomies are important factors in explaining item difficulty (Primi, 2014).

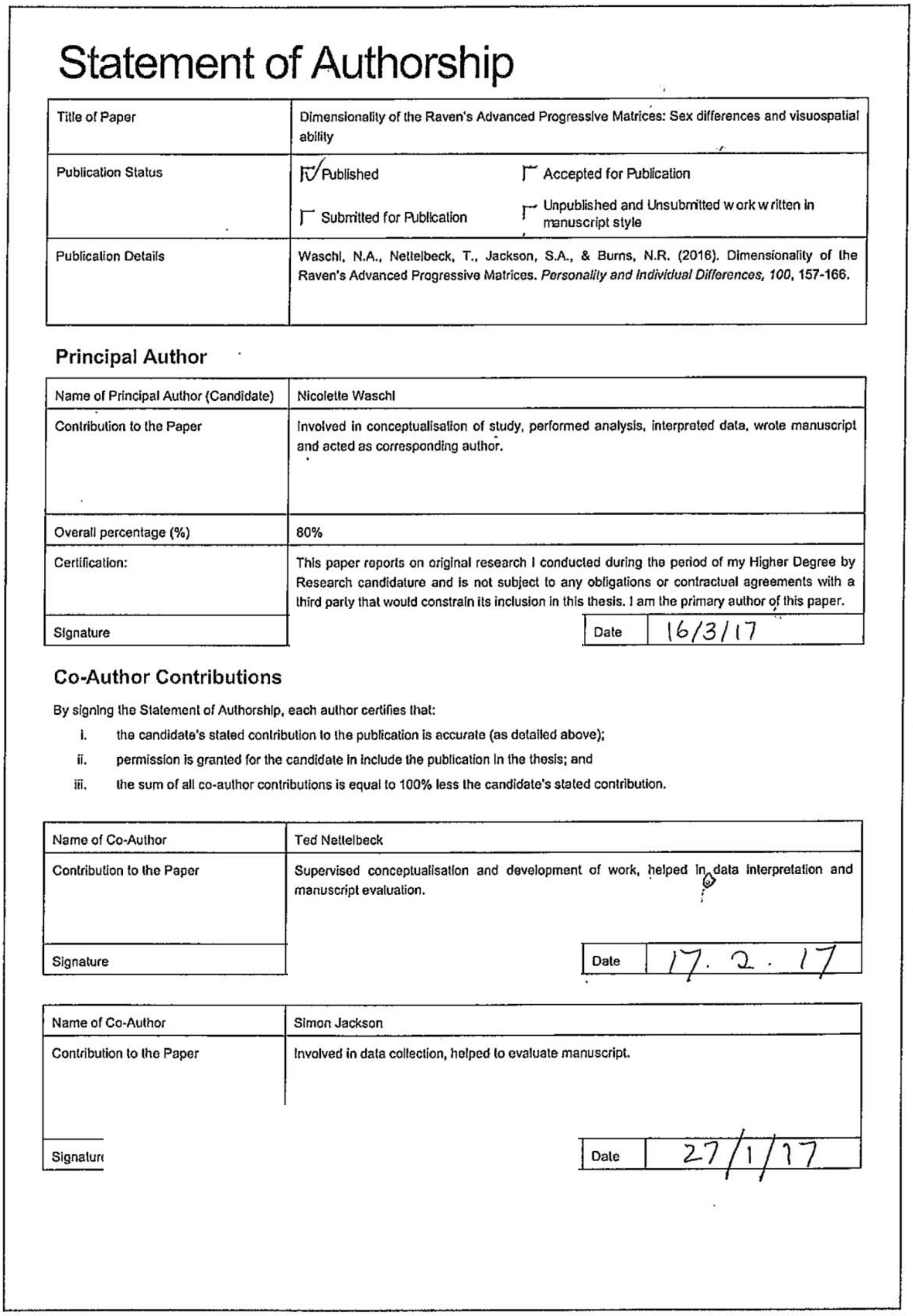
Following is Paper 1, which investigates the factor structure of three different versions of the APM, with a consideration of solution taxonomies, measurement invariance and differential item functioning.

# Paper 1

## Dimensionality of the Raven's Advanced Progressive Matrices: Sex differences and visuospatial ability

Waschl, N.A, Nettelbeck, T., Jackson, S. A., & Burns, N. R. (2016). Dimensionality of the Raven's Advanced Progressive Matrices: Sex differences and visuospatial ability. *Personality and Individual Differences, 100,* 157-166. doi:10.1016/j.paid.2015.12.008

*Note.* The published version of this paper and associated supplementary materials can be found in Appendix A.



# Statement of Authorship

| | |
|---|---|
| Title of Paper | Dimensionality of the Raven's Advanced Progressive Matrices: Sex differences and visuospatial ability |
| Publication Status | ☑ Published □ Accepted for Publication<br>□ Submitted for Publication □ Unpublished and Unsubmitted work written in manuscript style |
| Publication Details | Waschl, N.A., Nettelbeck, T., Jackson, S.A., & Burns, N.R. (2016). Dimensionality of the Raven's Advanced Progressive Matrices. *Personality and Individual Differences, 100*, 157-166. |

## Principal Author

| | |
|---|---|
| Name of Principal Author (Candidate) | Nicolette Waschl |
| Contribution to the Paper | Involved in conceptualisation of study, performed analysis, interpreted data, wrote manuscript and acted as corresponding author. |
| Overall percentage (%) | 80% |
| Certification: | This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper. |
| Signature | Date 16/3/17 |

## Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

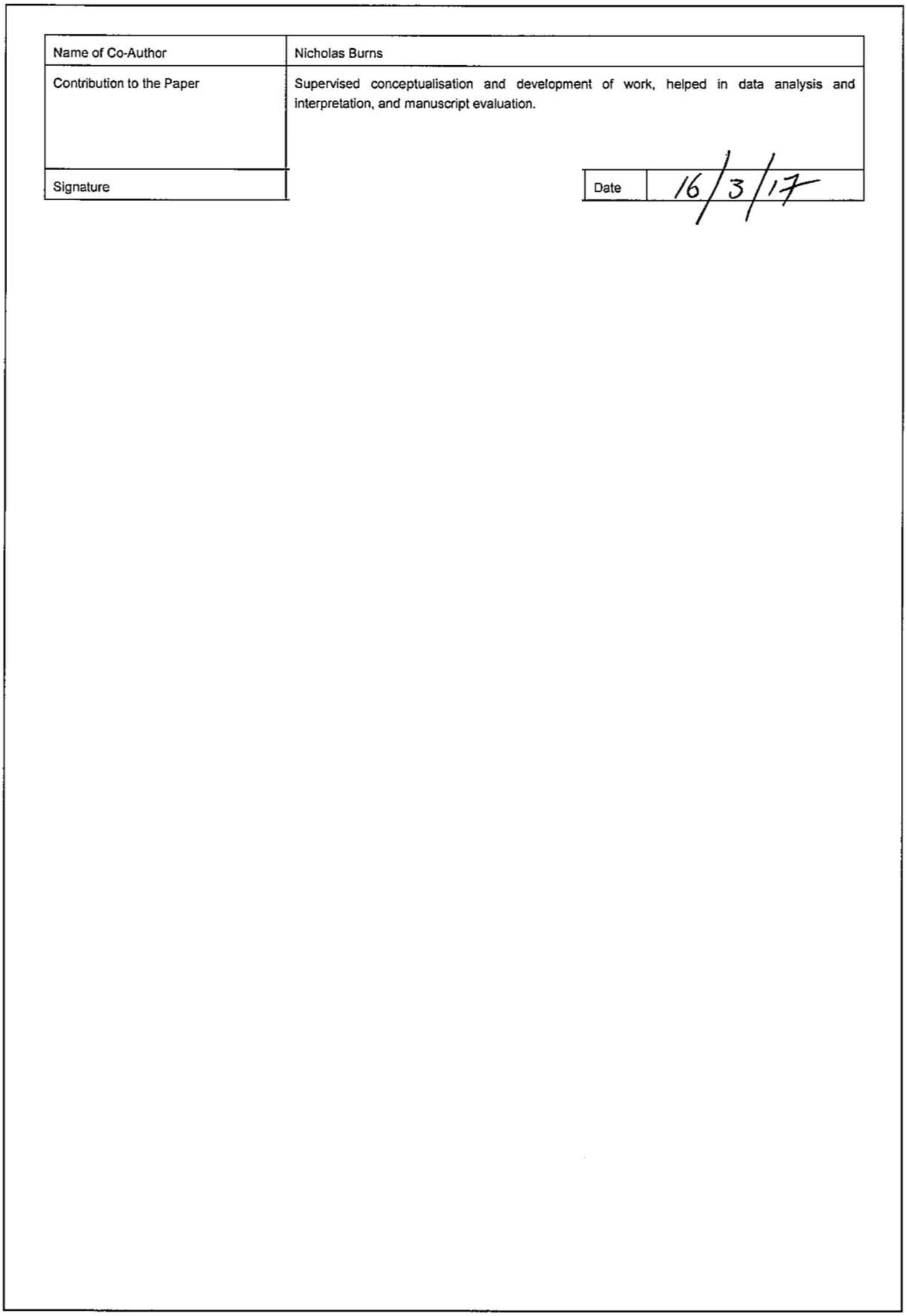
i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

| | |
|---|---|
| Name of Co-Author | Ted Nettelbeck |
| Contribution to the Paper | Supervised conceptualisation and development of work, helped in data interpretation and manuscript evaluation. |
| Signature | Date 17. 2. 17 |

| | |
|---|---|
| Name of Co-Author | Simon Jackson |
| Contribution to the Paper | Involved in data collection, helped to evaluate manuscript. |
| Signatur | Date 27/1/17 |



| | | | |
|---|---|---|---|
| Name of Co-Author | Nicholas Burns | | |
| Contribution to the Paper | Supervised conceptualisation and development of work, helped in data analysis and interpretation, and manuscript evaluation. | | |
| Signature | | Date | 16/3/17 |

## Abstract

Raven’s Progressive Matrices are considered a measure of inductive reasoning. However, there is evidence to suggest that they are not unidimensional, and they may measure visuospatial ability in addition to inductive reasoning. We investigated the psychometric properties of several versions of the Advanced Progressive Matrices (APM). Confirmatory factor analyses and Rasch analyses were used to investigate the dimensionality of the test, sex differences regarding dimensionality, and the utility of proposed taxonomies of item solution strategies. Three samples were administered three different forms of the test. Sample 1 consisted of 1297 individuals (929 females) who completed a 12-item short form; Sample 2 consisted of 455 individuals (327 females) who completed the full APM; and Sample 3 consisted of 362 individuals (244 females) who completed a 15-item short form. Results indicated that all three forms of the APM are unidimensional and measurement invariant across sex. There was little support for the validity of the taxonomies of solution strategies.

*Keywords*: Raven’s Progressive Matrices; Sex differences; Confirmatory factor analysis; Rasch analysis

Raven’s Progressive Matrices (RPM) were designed to measure Spearman’s *g*. Under the Cattell-Horn-Carroll model of intelligence (McGrew, 2009), the RPM tests (including the coloured, standard and advanced versions, designed for use with different populations) measure fluid intelligence and, specifically, inductive reasoning. However, there has been speculation that they also measure visuospatial ability (see Burke, 1958, for an early review). Fluid ability involves solving unfamiliar problems, while inductive reasoning, a narrow ability under fluid ability, involves discovering underlying principles or rules (McGrew, 2009). Visuospatial ability is different. It involves perceiving, generating and operating on visual patterns and stimuli, and is typified by tasks requiring perception and manipulation of visual forms (McGrew, 2009). It is clear why the claim that the RPM involves visuospatial ability emerged; RPM items comprise visual stimuli and it is conceivable that solving items could require visual transformation of these stimuli. This question was posed more than half a century ago, yet it remains unresolved. This paper focuses on the Advanced Progressive Matrices (APM).

The claim that the APM involves visuospatial ability has important implications. It is essential to understand what such a commonly used and potentially high-stakes test measures in order to understand how scores can be interpreted, used appropriately, and related to other constructs. Additionally, there is evidence of sex differences, favouring males, in APM performance (Lynn & Irwing, 2004a, 2004b). One of the most robust findings in the literature is a male advantage on visuospatial ability tests, particularly mental rotation (Linn & Petersen, 1985; Voyer, Voyer & Bryden, 1995). Therefore, one explanation, other than in terms of inductive reasoning, of a male advantage on the APM, could be the contribution of visuospatial ability to performance. Indeed, there is evidence that visuospatial ability accounts for the observed sex difference in APM scores (Colom, Escorial, & Rebollo, 2004). Through an understanding of whether the APM is unidimensional or multidimensional, and if this differs in male and females, we can come closer to understanding if and how visuospatial ability is involved.

Three different strategies have been used to understand what construct or constructs the APM measures: creation and examination of solution taxonomies based on information processing theories; investigation of sex differences in relation to these taxonomies; and factor analysis. This paper expands on these methods using three different versions of the APM.

Concerning solution taxonomies, Carpenter, Just, and Shell (1990) used patterns of eye fixations and verbalization of solution strategies to determine how each item was solved, resulting in a taxonomy of five solution rules (Table 1). These rules were: constant in a row; quantitative pairwise progression; addition/subtraction; distribution of three; and distribution of two. Constant in a row is not considered further because it always occurs in conjunction with another rule. Quantitative pairwise progression involves a quantitative increment or decrement across the row in size, position or number; Addition/Subtraction involves adding or subtracting a figure in one column from another figure to produce the third; Distribution of three is when three values from a categorical attribute are distributed across the row; and Distribution of two is when two values from a categorical attribute are distributed across a row and the third value is null.

Following Carpenter et al. (1990), DeShon, Chan, and Weissbein (1995) expanded on these rules and obtained 12 solution rules; four involved verbal-analytic processes and eight involved visual processes. Although these taxonomies are not directly comparable, Carpenter et al.'s addition/subtraction rule tended to equate with

DeShon et al.'s visual process, while distribution of three tended to equate with an analytic process.

Table 1

*Classifications of APM Items.*

| Item | Carpenter et al. (1990) | DeShon et al. (1995) | Dillon et al. (1981) | Item | Carpenter et al. (1990) | DeShon et al. (1995) | Dillon et al. (1981) |
|---|---|---|---|---|---|---|---|
| 1 | D3 | Analytic | | 19 | A/S | Both | |
| 2 | | Either | PP | 20 | A/S | Both | |
| 3 | P | Visual | PP | 21 | D3 | Analytic | A/S |
| 4 | P | Analytic | PP | 22 | D2 | Visual | |
| 5 | P | Either | PP | 23 | D2 | Visual | |
| 6 | P | Either | | 24 | P | Visual | |
| 7 | A/S | Visual | A/S | 25 | P | Both | |
| 8 | D3 | Analytic | | 26 | P, D3 | Both | PP |
| 9 | A/S | Visual | A/S | 27 | D3 | Analytic | |
| 10 | P | Visual | A/S | 28 | D3 | Analytic | A/S |
| 11 | | Visual | A/S | 29 | D3 | Analytic | |
| 12 | A/S | Visual | | 30 | D2 | Analytic | |
| 13 | D3 | Analytic | | 31 | D3, D2 | Both | |
| 14 | P | Either | | 32 | D3, D2 | Visual | |
| 15 | A/S | | | 33 | A/S | Visual | |
| 16 | A/S | Visual | A/S | 34 | D3 | Analytic | |
| 17 | D3 | Analytic | PP | 35 | D2 | Both | A/S |
| 18 | | Visual | | 36 | D2 | Analytic | PP |

*Note*. P = Quantitative Pairwise Progression; A/S = Addition/Subtraction; D3 = Distribution of 3; D2 = Distribution of 2; PP = Pattern Progression; Both = Analytic and Visual; Either = Analytic or Visual. Carpenter et al.'s (1990) classifications are supplemented by Mackintosh and Bennett (2005).

Given the well-established male advantage on visuospatial ability tests, and the grouping of solution rules into verbal-analytic and visual types, these solution taxonomies have been studied in relation to sex differences on the APM. Mackintosh and Bennett (2005) found a male advantage on items involving addition/subtraction and distribution of two, argued to involve visual processes, but no sex difference in items involving quantitative pairwise progression and distribution of three, argued to involve analytic processes. Other studies, however, have found no consistent sex differences in these item types (Vigneau & Bors, 2008), or a male advantage on all types (Colom & Abad, 2007). The picture of how these item types may relate to sex differences in scores is not clear.

Similarly, while factor analysis has commonly been used to investigate the structure of the APM, it has yet to provide a solution to the question of dimensionality, or the role of visuospatial ability in performance. One of the most cited factor analyses of the APM was by Dillon, Pohlmann, and Lohman (1981). Using a principal components analysis of phi/phi(max) coefficients, these authors reported two orthogonal factors, pattern progression and addition/subtraction (see Table 1). Addition/subtraction is broadly similar to Carpenter et al.'s (1990) addition/subtraction (although it is represented by different items in Dillon et al.'s study); while pattern progression involves perceiving a recurring or sequential design. However, later research has not supported Dillon et al.'s factors (Alderton & Larson, 1990; Arthur, Tubre, Paul, & Sanchez-Ku, 1999; Arthur & Woehr, 1993; Bors & Stokes, 1998; Vigneau & Bors, 2008) and other factor analyses have tended to indicate a single-factor structure (Abad, Colom, Rebollo, & Escorial, 2004; Chiesi, Ciancaleoni, Galli, & Primi, 2012; Schweizer, Goldhammer, Rauch, & Moosbrugger, 2007).

Both Carpenter et al. (1990) and DeShon et al.'s (1995) taxonomies have been used in factor analytic studies investigating the dimensionality of the APM. Unfortunately, although these rules have been useful in understanding the cognitive processing strategies that individuals use in solving individual items, there is little support for the idea that these rules represent different latent factors or relate to different latent abilities (Vigneau & Bors, 2008). One aspect yet to be investigated in relation to these solution taxonomies and sex differences in APM performance, however, is whether the latent structure of this test differs across sexes. There is some evidence that it may. For example, Lim (1994) found that the APM loaded on only one factor, formal operations, in males, but two, formal operations and spatial, in females. If this were the case, it could explain some of the inconsistent findings regarding the factor structure of the test. Relatedly, whether or not the test is measurement invariant across sex is important when considering the possibility of different factor structures among males and females, and when considering sex differences in the underlying construct. Despite consideration of the role of visuospatial ability and sex differences in APM performance, little concern has been given to establishing measurement invariance across sex in this test.

Although factor analytic studies have largely supported a unidimensional conceptualization of the APM, other lines of evidence indicate that the APM contains a visuospatial component or, at least, is not unidimensional. This evidence comes from studies using statistical control of visuospatial ability (Colom et al., 2004), experimental manipulation (DeShon et al., 1995), item response theory analysis (Vigneau & Bors, 2005) and neuroimaging (Ebisch et al., 2012); this uncertainty indicates that the matter deserves further consideration. The common finding of unidimensionality in the APM may be partially due to the various issues inherent in the use of factor analysis to answer this question. Ordinary factor analytic methods (e.g. principal axis factoring, maximum likelihood) applied to binary data can be problematic (Hattie, 1985). On the other hand, the weighted least squares mean and variance adjusted (WLSMV) estimator has several advantages over other methods. It was designed specifically for use with binary data and simulation studies have shown it to be appropriate for these types of data (Muthén, du Toit, & Spisic, 1997). However, this estimation method has not yet been applied to the full form APM or the two short forms considered in the present study (Bors & Stokes, 1998; and a form unique to this study).

Another method for investigating the dimensionality of the APM utilizes item response theory (IRT), which is not subject to the same issues as factor analytic methods. Unlike factor analysis, IRT was created for binary data and is therefore appropriate for use with the data obtained from the correct-incorrect responses to APM items. While factor analysis conducted using the WLSMV estimator and IRT are mathematically highly similar, their distinct theoretical standpoints provide an interesting comparison. There are several different IRT models, including the Rasch, 2PL and 3PL models. The Rasch model considers the probability of a correct response to an item given the test-taker’s ability and the item difficulty while holding constant item discrimination and guessing. The 2PL and 3PL models allow estimation of other parameters in addition to difficulty; the 2PL model allows estimation of item discrimination, while the 3PL model allows estimation of discrimination and guessing. The Rasch model has excellent measurement properties and well-developed statistical theory, and hence has been used here.

While some studies have used IRT to investigate the APM, few have applied the Rasch model, and those that have did not consider sex differential item functioning (DIF; Vigneau & Bors, 2005). DIF has been considered under the 2PL (Abad et al.,

2004) and 3PL models (Chiesi, Ciancaleoni, Galli, Morsanyi, & Primi, 2012; Chiesi, Ciancaleoni, Galli, & Primi, 2012), with conflicting findings. Using the 2PL model, Abad et al. (2004) showed more DIF for items classified as visuospatial than items classified as analytic, while Chiesi, Ciancaleoni, Galli, Morsanyi, et al. (2012) and Chiesi, Ciancaleoni, Galli, and Primi's (2012) work indicated no DIF.

Hence, while there has been a significant amount of work conducted on whether the APM is unidimensional or whether it involves a second ability, hypothesised to be visuospatial ability, questions still remain. The aims of this study were threefold. First, to investigate sex differences in individual items and item types in an attempt to clarify the existing literature and as a prelude to investigating sex differences in the latent structure of the test. Secondly, we applied factor analytic methods not yet used on the APM to test models based on results of Carpenter et al. (1990), DeShon et al. (1995) and Dillon et al. (1981) in males and females separately, and, if appropriate, to examine measurement invariance and latent mean differences. Thirdly, we examined Rasch model fit and DIF to supplement the factor analytic results.

# Method

## Participants and Measures

All participants provided informed consent before participating in these studies. Participants in Sample 1 were 1297 individuals tested through the University of Adelaide, Australia, most of whom were university students. Participants completed a 12-item short form of the APM (Bors & Stokes, 1998) online and in their own time as part of their coursework. The items originated from Set II of the APM, and were included based on high item-total correlations and low inter-item correlations. The final pool of items consisted of items 3, 10, 12, 15, 16, 18, 21, 22, 28, 30, 31 and 34 from the full form. Nine cases with a score of zero were deleted from this dataset because it was presumed that these participants did not understand the instructions or had not taken the test seriously. The final sample consisted of 1288 (929 females) aged 16 to 60 years (*M* = 23.5, *SD* = 6.62). The mean score for this sample was 7.19 items (60% correct; *SD* = 2.71). Males (*M* = 7.53 [63%], *SD* = 2.68) scored slightly but significantly higher than females (*M* = 7.06 [59%], *SD* = 2.71), *t*(1286) = 2.78, *p* = .005, *d* = .17.

Participants in Sample 2 were 455 adults (327 females) aged 16 to 68 years (*M* = 34.47, *SD* = 16.9) residing in Adelaide and recruited over two studies (see Burns, Bastian, & Nettelbeck, 2007). Each participant completed a paper-and-pencil version of Set II of the full form of the APM (36 items; Raven, 1962). The mean score for Sample 2 was 21.34 (59%; *SD* = 7.24). Males (*M* = 22.88 [64%], *SD* = 6.79) scored significantly higher than females (*M* = 20.74 [58%], *SD* = 7.34), *t*(453) = 2.845, *p* = .005, *d* = .30.

Participants in Sample 3 were 362 undergraduate psychology students from the University of Sydney, recruited over two studies (Jackson, Kleitman, Stankov, & Howie, n.d.). Participants completed a 15-item short form of the APM, unique to these studies, as part of a larger test battery either online in their own time or as part of their regular tutorial programme. The items originated from Set II of the APM, and the criteria for inclusion of items was based on pilot testing designed to obtain a more pure measure of the APM by selecting those items showing high item-total and inter-item correlations, and high standard deviations. The final pool of items consisted of items 7, 11, 13, 15, 16, 17, 18, 21, 23, 25, 26, 27, 30, 32 and 34 from the full form. Two cases with a score of zero were removed from this dataset and the final sample consisted of 360 (244 females) aged 17 to 54 years (*M* = 20.09, *SD* = 3.19). The mean score for

Sample 3 was 8.39 (56%; *SD* = 3.42). Males (*M* = 9.02 [60%], *SD* = 3.70) scored significantly higher than females (*M* = 8.09 [54%], *SD* = 3.24), t(358) = 2.41, *p* = .016, *d* = .28. The sex difference on this short form was substantially larger than the difference in Sample 1, but did show a similar effect size to Sample 2. The method of item selection for the two short forms was different, resulting in only six (40-50%) common items between these short forms, which may have influenced the magnitude of the differences.

### Data Analysis

**Sex differences in items and item types**. Analysis of sex differences in individual items and groups of item types was conducted in R (R Core Team, 2014). Chi-square tests were used to examine sex differences in individual items and t-tests were used to investigate sex differences in groups of item types as classified by Carpenter et al. (1990), DeShon et al. (1995) and Dillon et al. (1981).

**Confirmatory factor analysis.** Confirmatory factor analysis was performed in Mplus 7 (Muthén & Muthén, 1998-2012) using the WLSMV estimator. Several models were compared: models based on the item classifications of Carpenter et al. (1990), DeShon et al. (1995) and Dillon et al. (1981), and a model based on item threshold values (difficulty model)[1]. These models were compared with reference to the chi-square value, RMSEA and CFI. Guidelines for interpreting these indices recommend

[1] Bi-factor models were also considered, however these models presented estimation problems and demonstrated poor fit when successfully estimated. Therefore this type of model was deemed too complex for the data.

the following cut-off values for acceptable fit: normed chi-square (i.e. $\chi^2$/df) < 2 (Kline, 1998), RMSEA < .05 (Browne & Cudeck, 1993) and CFI > .95 (Hu & Bentler, 1999). In addition to inspection of fit indices, where possible, the multi-factor models were compared to their corresponding one-factor models by comparing the chi-square statistics. Because the WLSMV estimator does not follow the chi-square distribution, the DIFFTEST function (Asparaouhov & Muthén, 2006) in Mplus was used to test for differences between the models. This function follows a two-step process: first, the less restrictive model is estimated and the derivatives needed for the chi-square difference test are saved. Secondly, the more restrictive model is estimated and the chi-square difference test is calculated using the derivatives from both models (Muthén & Muthén, 1998-2012). Given that not all items were always included in the multi-factor models because some items were not classified under the relevant taxonomies, the corresponding one-factor models only included those items that were in the multi-factor model.

**Measurement invariance and latent mean differences**. If factor analysis demonstrated that the best fitting model was the same for males and females, multiple-groups confirmatory factor analysis (MGCFA) was carried out in order to determine, firstly, if measurement invariance across sex could be confirmed, and secondly, if there were any differences in the latent mean (or means if considering a multi-factor model) of APM performance. Measurement invariance was tested using mean and covariance structures (MACS) with delta parameterization in Mplus 7.

MGCFA for categorical indicators is somewhat different from MGCFA with continuous indicators. When the data consist of continuous indicators, increasingly restrictive models are tested by constraining equal the factor loadings, intercepts, and then residual variances of factor indicators. With categorical indicators, MGCFA involves testing factor loadings, thresholds and scale factors. Thresholds are tested instead of intercepts and scale factors may be tested instead of residual variances. Categorical MGCFA also involves a comparison of only two models, a less restrictive and a more restrictive, rather than the usual four when dealing with continuous data. The first, less restrictive model allows the thresholds and factor loadings to vary across groups, while scale factors are constrained at one and factor means are constrained at zero in all groups. This less restrictive model is then compared to the more restrictive model in which thresholds and factor loadings are simultaneously held equal across groups and the scale factor is fixed to one and the factor means constrained at zero in the first group but allowed to vary in the other. If there is a significant difference in the fit between these models, this indicates a violation of measurement invariance. To test for the difference in fit, the DIFFTEST function, as explained in section 2.2.2, was used. If there was no significant difference in model fit, invariance was met. If a significant difference in model fit was found, the modification indices were inspected to determine which item may be responsible, and the model was re-tested allowing the factor loading and threshold to vary across groups, and constraining that scale factor to one across groups. Once partial invariance was established (i.e. the problem item's measurement parameters were allowed to vary while all others were constrained equal), examination of latent mean differences was conducted by comparing the latent mean in the non-reference group to that of the reference group, which was constrained at zero. A significant value indicated a significant difference in the latent mean across groups (i.e., across sex).

**Rasch analysis and differential item functioning.** Rasch analysis was conducted using ConQuest 3.0.1 (Adams, Wu, & Wilson, 2012). In order to examine whether the data conformed to a Rasch model, and therefore could be considered unidimensional, person fit and item fit were inspected using the mean square statistics. The infit mean square considers the consistency of the item responses to the item characteristic curve (ICC) for each item, with weighted consideration of the responses of those persons close to the 0.5 probability level for that item. Low infit mean square values indicate item redundancy, while high values are more of a threat to unidimensionality because they indicate that the item discriminates poorly. The criterion for person (case) misfit was a standardized outfit mean square > 5, represented by the *t*-value. A value exceeding this cutoff indicates an erratic response pattern. The criterion for item misfit was a standardized infit mean square (*t*-value) > 2. While generally the critical value for the unstandardized infit mean square is in the range of 0.77 – 1.30 (Adams & Khoo, 1993) in large samples a *t*-value > 2 can be within this range. The *t*-value allows analysis of strict conformity to a Rasch model (perfect model fit) as opposed to whether or not the items are productive for measurement, for which the unstandardized values are more useful (Linacre, 2002). Therefore, for the purposes of investigating the dimensionality of these measures, it was decided that a stricter rather than a more lenient assessment of fit was appropriate and *t*-values were applied.

In addition to assessment of dimensionality using item and person fit, a principal components analysis (PCA) of the Rasch residuals (residual variance in the data once the variance explained by the Rasch dimension is accounted for) was performed. If the first principal component of the Rasch residuals is above noise level, this indicates the presence of multidimensionality. A simulation based on the number of items and number of cases was performed for each sample to determine at what point the value of the first eigenvalue exceeded random noise (Raîche, 2005). If this value was exceeded, this indicated multidimensionality.

Following examination of dimensionality, DIF across sex was investigated using two methods; the item fit and the item threshold approaches. The item fit approach involved calibration of the Rasch model separately in the two groups and inspection of the item fit *t*-values in order to determine if the same items showed misfit. If an item displays acceptable fit in the combined group, but shows misfit in only one of the male or female groups, this indicates a biased item that does not discriminate equally across groups (Hungi, 2005). The item threshold approach used the Wald *t* statistic, calculated on the basis of the values provided by the item-by-sex interaction parameters in ConQuest. The Wald *t* statistic is calculated by dividing the item's item-by-group interaction parameter by its standard error. Any item with a *t*-value > 2 displays statistically significant DIF.

## Results

### Sex Differences in Items and Item Types

In each sample there were individual items showing significant sex differences, all of which showed a higher male score (Table 2). Item numbers are presented as their number in the original form of the test (i.e. the first item in Sample 1 is labeled as item 3) so as to facilitate comparison across forms.

No items were found to show a consistent sex difference across samples, and there was no clear pattern regarding the classifications of the individual items showing sex differences. Similarly, no consistent pattern of sex differences in item types was found (Table 3). In Samples 1 and 2 nearly all groups of items showed a significant sex difference, while in Sample 3 only distribution of two (Carpenter et al., 1990), visual

(DeShon et al., 1995) and addition/subtraction (Dillon et al., 1981) showed significant differences.

Table 2

*Sex differences in individual items.*

| Test | Item | Carpenter et al. (1990) | DeShon et al. (1995) | Dillon et al. (1981) | Chi-square | Phi |
|---|---|---|---|---|---|---|
| Sample 1 | 12 | A/S | Visual | - | 10.36** | .09 |
| | 21 | D3 | Analytic | A/S | 5.61* | .07 |
| | 22 | D2 | Visual | - | 7.21 ** | .08 |
| | 28 | D3 | Analytic | A/S | 4.13* | .06 |
| Sample 2 | 2 | - | Either | PP | 6.32* | .12 |
| | 4 | P | Analytic | PP | 10.02** | .15 |
| | 9 | A/S | Visual | A/S | 7.41** | .13 |
| | 10 | P | Visual | A/S | 4.46* | .10 |
| | 13 | D3 | Analytic | - | 5.48* | .11 |
| | 21 | D3 | Analytic | A/S | 4.49* | .10 |
| | 22 | D2 | Visual | - | 5.71* | .11 |
| | 25 | P | Both | - | 6.95** | .12 |
| | 26 | P, D3 | Both | PP | 5.49* | .11 |
| Sample 3 | 30 | D2 | Analytic | - | 4.08* | .11 |
| | 32 | D3, D2 | Visual | - | 9.55** | .16 |

*Note*. P = Quantitative Pairwise Progression; A/S = Addition/Subtraction; D2 = Distribution of two; D3 = Distribution of three; PP = Pattern Progression; Both = Analytic and Visual; Either = Analytic or Visual. All significant differences favour males.

* *p* < .05; ** *p* < .01

Table 3

*Sex differences in item types.*

| Taxonomy | Classification | *t* | | |
|---|---|---|---|---|
| | | Sample 1 | Sample 2 | Sample 3 |
| Carpenter et al. (1990) | Pairwise progression | 1.25 | 3.11** | |
| | Addition/Subtraction | 2.77**[a] | 2.05* | 1.04[b] |
| | Distribution of 3 | 2.51* | 2.43* | 1.29 |
| | Distribution of 2 | 2.28* | 2.35* | 2.55* |
| DeShon et al. (1995) | Visual | 2.30* | 2.49* | 2.69** |
| | Analytic | 2.33* | 2.35* | 1.69 |
| Dillon et al. (1981) | Pattern Progression | | 3.60** | 0.71[c] |
| | Addition/Subtraction | 2.59** | 2.77** | 2.01* |

*Note.* Sample 1 *df* = 1286; Sample 2 *df* = 453; and Sample 3 *df* = 358. [a]*df* = 732.47 [b]*df* = 283.70 [c]*df* = 295.54. All significant differences favour males.
Blank cells indicate groups that could not be tested because there were not enough items to represent the classification
* $p < .05$; ** $p < .01$

## Confirmatory Factor Analysis

This section presents the results of the factor analysis in each sample (see supplementary materials for additional information). In all cases, the difficulty model showed an acceptable fit to the data and was statistically significantly better fitting than the corresponding one-factor model (with the exception of the Sample 3 combined and female data). One common issue with the use of factor analysis with the type of data used in the current study is the occurrence of artifactual factors based on item difficulty. The use of the WLSMV estimator should avoid the issue of these artifactual difficulty factors by allowing a non-linear relationship between the item and the factor, the main cause of this problem (Gibson, 1960). However, the difficulty models are statistically derived models, in that item classifications of ‘easy’ and ‘hard’ were determined from item threshold values, rather than derived theoretically. Therefore, the difficulty models are presented here for completeness, but will not be discussed further in this section. The interpretation of these difficulty models will be discussed in Section 4, below.

There were several appearances of Heywood cases in some multi-factor models. This is caused by negative error variance and/or a correlation between two factors exceeding 1. One cause is the estimation of too many factors and it is considered likely that this was the cause of the appearance of these cases here. Hence, the models where this occurred were considered invalid.

**Sample 1.** The results from the factor analysis of Sample 1 data are presented in Table 4. In this dataset, the model based on Dillon et al. (1981) could not be calculated because there were not enough items to represent the pairwise progression factor. The results obtained for the analysis of the combined dataset show that the one-factor and DeShon et al. (1995) models had similar and acceptable fit indices. The DeShon et al. model fit statistically significantly better than the one-factor model. However, the correlation between these two factors was too high (> .90) for two separate factors to be considered meaningful. Therefore, it can be concluded that the one-factor model provided the best fit for the combined group.

Table 4

*Model fit indices: Sample 1.*

| Model | $\chi^2$ | *df* | CFI | RMSEA | Factor correlation | |
|---|---|---|---|---|---|---|
| | | | | | Estimate | 95% CI |
| One-Factor | | | | | | |
| Combined | 103.84 | 54 | .98 | .03 | | |
| Female | 84.60 | 54 | .98 | .03 | | |
| Male | 80.32 | 54 | .97 | .04 | | |
| Two factors: DeShon et al. (1995) | | | | | | |
| Combined* | 58.72 | 34 | .99 | .02 | .91 | .85 – .97 |
| Female* | 49.07 | 34 | .99 | .02 | .88 | .79 – .97 |
| Male | 52.42 | 34 | .97 | .04 | .98 | .87 – 1.11 |
| Four factors: Carpenter et al. (1990) | | | | | | |
| Combined | - | | | | | |
| Female* | 54.91 | 38 | .99 | .02 | .69 - .95 | .51 – 1.12 |
| Male | - | | | | | |
| Two factors: Difficulty | | | | | | |
| Combined* | 72.77 | 53 | .99 | .02 | .82 | .75 – .88 |
| Female* | 68.38 | 53 | .99 | .02 | .84 | .75 – .91 |
| Male* | 64.25 | 53 | .99 | .02 | .77 | .64 – .89 |

* These models were significantly better fitting than their corresponding one-factor models. DeShon et al. (combined) **Δ** $\chi^2$ (1) = 5.87, *p* = .015; DeShon et al. (female) **Δ** $\chi^2$ (1) = 7.02, *p* = .008; Carpenter et al. (female) **Δ** $\chi^2$ (6) = 21.00, *p* = .002; difficulty (combined) **Δ** $\chi^2$ (1) = 24.76, *p* < .001; difficulty (female) **Δ** $\chi^2$ (1) = 13.99, *p* < .001; and difficulty (male) **Δ** $\chi^2$ (1) = 12.03, *p* < .001

Blank rows indicate models could not be calculated due to the presence of Heywood cases

The comparison of the results obtained from the female and the male data give some additional insight. The male data tended to fit the one-factor model best. However, in the female group, the multi-factor models were consistently better fitting than their corresponding one-factor models, despite a high factor correlation. These results indicate that there could be a sex difference in the latent structure of this test.

**Sample 2**. Table 5 presents the results from Sample 2. In calculating these models it became apparent that several pairs of items showed an empty cell in the bivariate table – that is, were statistically indistinguishable – especially in the male data. To deal with this, items from problem pairs were systematically deleted. This involved the deletion of those items that were present in the largest number of problematic pairs until there were no longer any problematic pairs left. Where this left multiple combinations, the items retained were those that resulted in the best model fit. Item 36 was not included in these models because it was almost always present in these problem pairs, and only a small number of participants successfully answered it. There were several other problem pairs, which tended to involve items 4, 9, 24, 29 and 35. No consistent characteristic of these items was found to explain this. Therefore, for Sample 2 models, the items used for male data and the female data are not directly comparable.

Overall, the results indicated that, given the high correlations between factors in the multi-factor models, a one-factor model best represented the data in all groups. Unlike the Sample 1 data, the female models tended to have factors too highly correlated (.92 and .94) for the fit of the multi-factor model to be considered acceptable, although the DeShon et al. (1995) model did fit statistically significantly better than its corresponding one-factor model. However, there was a trend for the female models to have less highly correlated factors than the male models.

Table 5

*Model fit indices: Sample 2.*

| Model | $\chi^2$ | *df* | CFI | RMSEA | Factor Correlation | |
|---|---|---|---|---|---|---|
| | | | | | Estimate | 95% CI |
| One-Factor | | | | | | |
| Combined | 755.22 | 560 | .97 | .03 | | |
| Female | 662.90 | 560 | .98 | .02 | | |
| Male | 462.64 | 434 | .97 | .02 | | |
| Two factors: DeShon et al. (1995) | | | | | | |
| Combined | 407.36 | 251 | .96 | .04 | .96 | .92 – 1.00 |
| Female* | 344.68 | 251 | .96 | .03 | .94 | .87 – .99 |
| Male | 221.06 | 188 | .96 | .04 | .99 | .91 – 1.06 |
| Two factors: Dillon et al. (1981) | | | | | | |
| Combined | 89.07 | 76 | .99 | .02 | .95 | .86 – 1.02 |
| Female | 83.54 | 76 | .99 | .02 | .92 | .82 – 1.01 |
| Male | - | | | | | |
| Two factors: Difficulty | | | | | | |
| Combined* | 693.36 | 559 | .98 | .02 | .86 | .78 – .89 |
| Female* | 619.10 | 559 | .99 | .02 | .85 | .77 – .90 |
| Male* | 457.13 | 433 | .98 | .02 | .89 | .76 – .96 |

* These models were significantly better fitting than their corresponding one-factor models. DeShon et al. (female) $\Delta\ \chi^2$ (1) = 4.93, *p* = .027; difficulty (combined) $\Delta\ \chi^2$ (1) = 34.70, *p* < .001; difficulty (female) $\Delta\ \chi^2$ (1) = 24.72, *p* < .001; difficulty (male) $\Delta\ \chi^2$ (1) = 5.91, *p* = .015

The model based on Carpenter et al. (1990) could not be calculated due to the presence of several Heywood cases across all groups

**Sample 3.** Sample 3 results are presented in Table 6. Similar to Sample 2, there were several pairs of items in the male data set that were indistinguishable. All pairs of statistically indistinguishable items involved items 7 and 11; hence these items were excluded from the models in the male group. Again, no characteristic of these items, which they did not share with many other items that were not problematic, was found to explain this. The results show that a one-factor model provided the best fit to the data, in all groups. Although the DeShon et al. (1995) model also fit the data in the combined and female groups, the factor correlations were again high and, unlike in the other samples, this model did not fit significantly better than the corresponding one-factor model.

Table 6

*Model fit indices: Sample 3.*

| Model | $\chi^2$ | *df* | CFI | RMSEA | Factor Correlation | |
|---|---|---|---|---|---|---|
| | | | | | Estimate | 95% CI |
| One-Factor | | | | | | |
| Combined | 133.25 | 90 | .97 | .04 | | |
| Female | 118.90 | 90 | .96 | .04 | | |
| Male | 78.34 | 65 | .98 | .04 | | |
| Two factors: DeShon et al. (1995) | | | | | | |
| Combined | 78.56 | 53 | .98 | .04 | .92 | .81 – 1.02 |
| Female | 76.54 | 53 | .96 | .04 | .93 | .78 – 1.06 |
| Male | 49.22 | 34 | .97 | .06 | .95 | .81 – 1.08 |
| Two factors: Dillon et al. (1981) | | | | | | |
| Combined | - | | | | | |
| Female | 19.13 | 8 | .93 | .08 | .96 | .68 – 1.52 |
| Male | - | | | | | |
| Two factors: Difficulty | | | | | | |
| Combined | 131.12 | 89 | .97 | .04 | .94 | .85 – 1.01 |
| Female | 115.03 | 89 | .96 | .04 | .89 | .77 – .98 |
| Male* | 72.02 | 64 | .99 | .03 | .88 | .74 – .96 |

* These models were significantly better fitting than their corresponding one-factor models. Difficulty (male) $\mathbf{\Delta}\ \chi^2$ (1) = 5.66, *p* = .017

The model based on Carpenter et al. (1990) could not be calculated due to the presence of several Heywood cases in all groups

## Measurement Invariance and Latent Mean Differences

**Sample 1.** Although there was some evidence in this sample that the female group fit a multi-factor model, the fit of the female data to the one-factor model was also acceptable, and hence measurement invariance testing was carried out using the one-factor model. Partial measurement invariance was met by allowing the factor loading and threshold of item 3 to vary across groups while constraining equal all other parameters (see Table 7 for relevant statistics). Item 3 showed a higher loading in the male group. The latent mean in the male group was significantly higher than that of the female group (0.198, *p* = .013).

**Sample 2.** The models were computed excluding items 4, 9, 29, 35 and 36 due to the presence of empty cells in the bivariate table in the male data. Partial measurement invariance was met by allowing the factor loading and threshold of item 8 to vary across groups while constraining equal all other parameters (see Table 7). Item 8 showed a higher loading in the male group. The latent mean in the male group was significantly higher than that of the female group (0.270, *p* = .028).

**Sample 3.** The models were computed excluding items 7 and 11 due to empty cells in the bivariate table in the male data. Partial measurement invariance was met by allowing the factor loading and threshold of item 32 to vary across groups while constraining equal all other parameters (see Table 7). Item 32 showed a higher loading in the male group. The latent mean in the male group was significantly higher than that of the female group (0.295, *p* = .020).

Table 7

*Measurement invariance statistics.*

| | χ2 | *df* | Δ χ2 | Δ *df* | Δ *p* | RMSEA | CFI |
|---|---|---|---|---|---|---|---|
| Sample 1 | | | | | | | |
| Less Restrictive | 164.79 | 108 | | | | .03 | .98 |
| More Restrictive | 192.18 | 118 | 24.54 | 10 | .006 | .03 | .97 |
| More Restrictive (Partial) | 178.39 | 117 | 14.66 | 9 | .101 | .03 | .98 |
| Sample 2 | | | | | | | |
| Less Restrictive | 975.58 | 868 | | | | .02 | .98 |
| More Restrictive | 1016.88 | 897 | 46.05 | 29 | .023 | .02 | .97 |
| More Restrictive (Partial) | 1008.28 | 896 | 37.47 | 28 | .109 | .02 | .98 |
| Sample 3 | | | | | | | |
| Less Restrictive | 165.58 | 130 | | | | .04 | .97 |
| More Restrictive | 196.17 | 141 | 28.90 | 11 | .002 | .05 | .96 |
| More Restrictive (Partial) | 170.61 | 140 | 7.17 | 10 | .709 | .04 | .98 |

**Rasch Analysis and Differential Item Functioning**

This section presents the results of the Rasch analysis for all samples. Figure 1 displays the test information curves (excluding items showing poor fit; see below) for each test version. There was little difference in the test information curves for the combined, female and male groups in their respective samples. Therefore, only the curves pertaining to the combined group are displayed. As expected, the full form used in Sample 2 provided the greatest information overall.

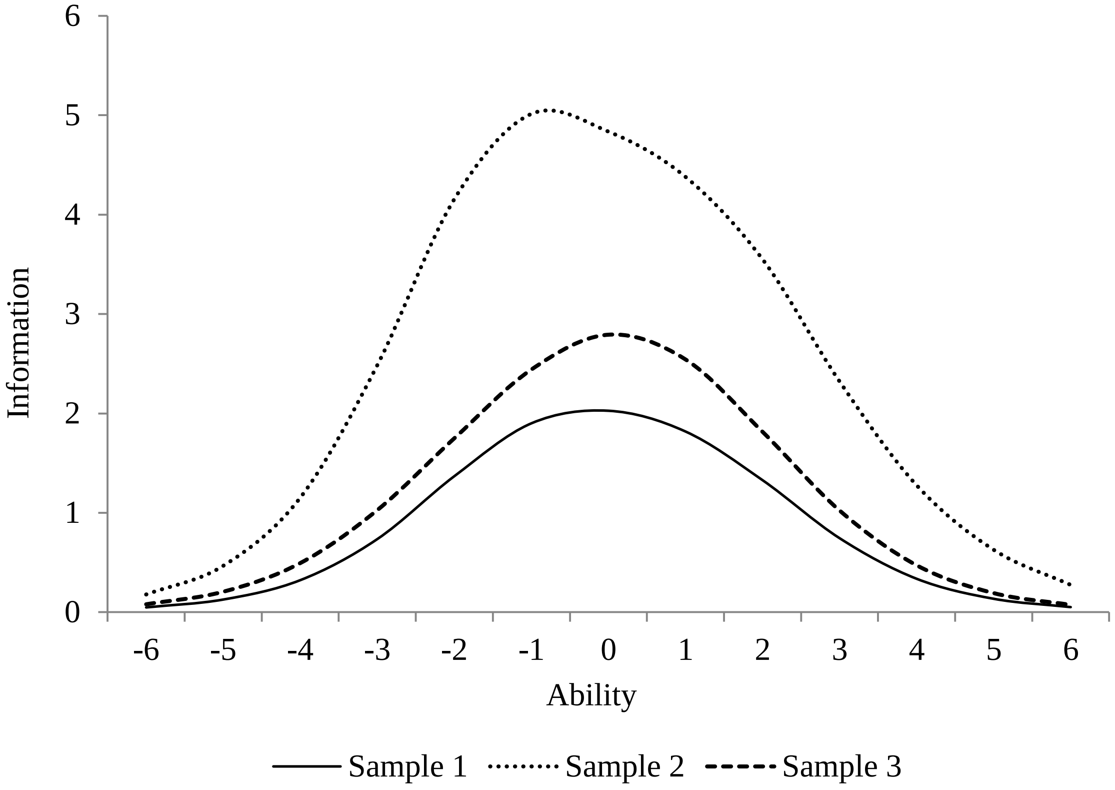


*Figure 1.* Test information curves for the combined group for each sample.

**Sample 1.** One case showed misfit (*t* = 6.43) and was excluded from analysis. Examination of item fit indices resulted in the exclusion of one item (item 21 [infit = 0.95, *t* = -2.2]). Given the negative t-value, this item was redundant, and therefore less of a threat to unidimensionality. There were no further misfitting items in the male or female groups when calibrated separately.

The Rasch dimension explained 36% of the variance. The first eigenvalue from the PCA of Rasch residuals was 1.09, explaining 9.9% of the residual variance. This was below the cut-off value of 1.16, supporting unidimensionality of the measure.

Although the variance explained by the Rasch dimension was lower than desirable, there was little evidence of a substantial second dimension. The female and male data also showed no evidence of multidimensionality (eigenvalues of 1.11 and 1.19 respectively, with cut-off values of 1.19 and 1.31). Figure 2 shows the ICCs for three representative items (easy, medium and difficult) in Sample 1 for the combined group.

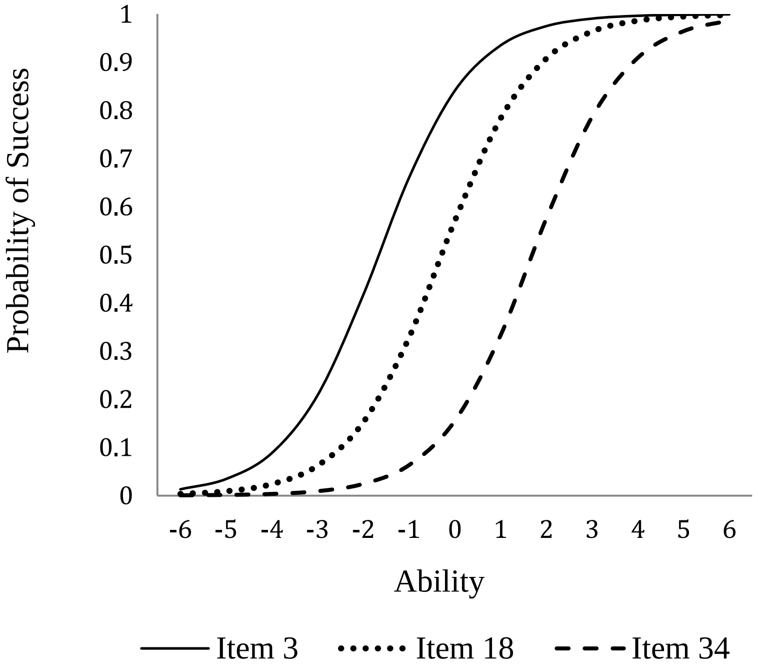


*Figure 2.* ICCs for three representative items from Sample 1 (combined group).

Significant DIF was found in two items; Item 12 was easier for males, while item 18 was easier for females. The column labeled sex x item (Table 8) displays the interaction term between sex and item, and displays the female difficulty estimate, where the mid-point between female and male difficulty levels is zero. For example, the value of -0.049 for item 3 indicates that this item was easier for females, with a difference of 2 x 0.049 (0.099) logits between the difficulty for males and females.

Table 8

*Item fit and DIF: Sample 1.*

| Item | Parameter<br>Estimate | SE | Infit | *t* | Sex<br>x item | SE | Wald *t* |
|---|---|---|---|---|---|---|---|
| 3 | -1.665 | .084 | 1.02 | 0.3 | -0.049 | .095 | -0.52 |
| 10 | -1.266 | .076 | 0.98 | -0.4 | 0.018 | .087 | 0.21 |
| 12 | -0.824 | .069 | 0.98 | -0.6 | 0.198 | .082 | 2.41* |
| 15 | -1.185 | .075 | 1.03 | 0.7 | 0.038 | .086 | 0.44 |
| 16 | -0.851 | .070 | 0.97 | -0.8 | -0.073 | .078 | -0.94 |
| 18 | -0.284 | .064 | 1.01 | 0.4 | -0.160 | .071 | -2.25* |
| 21 | | | | | | | |
| 22 | 0.546 | .061 | 0.96 | -1.7 | 0.101 | .069 | 1.46 |
| 28 | 1.785 | .068 | 1.03 | 0.9 | 0.063 | .073 | 0.86 |
| 30 | 1.053 | .062 | 1.01 | 0.3 | -0.052 | .069 | 0.75 |
| 31 | 0.993 | .062 | 1.02 | 0.7 | -0.015 | .069 | -0.22 |
| 34 | 1.696 | .067 | 1.00 | 0.1 | -0.068 | .073 | -0.93 |

* Indicates the presence of significant DIF at *p* < .05

**Sample 2.** Two misfitting cases were identified (*t* = 10.45 and 6.64) and excluded from analysis. Several misfitting items were identified and iteratively excluded from analysis, resulting in the exclusion of five (14%) items. Table 9 displays these items, their infit mean square and corresponding t-values, and their rule classification. Four items showed poor discrimination, while one showed redundancy (item 21). Although most of these items were classified as analytic, they represented only 33% of the total analytic items, indicating that this characteristic is unlikely to be responsible for this finding.

The Rasch dimension explained 42% of the variance. The first eigenvalue from the PCA of Rasch residuals was 1.36, explaining 3.9% of the residual variance. This was below the cut-off value of 1.58, supporting unidimensionality of the measure. Although the variance explained by the Rasch dimension was again lower than desirable, there was little evidence of a substantial second dimension. The female and male data also showed no evidence of multidimensionality (eigenvalues of 1.45 and 1.58 respectively, with cut-off values of 1.69 and 2.20). Figure 3 shows the ICCs for three representative items in Sample 2 for the combined group.

Table 9

*Misfitting items: Sample 2.*

| Item | Infit *(95% CI)* | *t* | Carpenter et al. (1990) | DeShon et al. (1995) | Dillon et al. (1981) |
|---|---|---|---|---|---|
| 13 | 1.15 *(0.91 – 1.09)* | 3.2 | D3 | Analytic | - |
| 21 | 0.87 *(0.91 – 1.09)* | -3.0 | D3 | Analytic | A/S |
| 17 | 1.18 *(0.88 – 1.12)* | 2.7 | D3 | Analytic | PP |
| 28 | 1.14 *(0.88 – 1.12)* | 2.3 | D3 | Analytic | A/S |
| 20 | 1.13 *(0.89 – 1.11)* | 2.2 | A/S | Both | - |

*Note*. Items displayed in order of exclusion

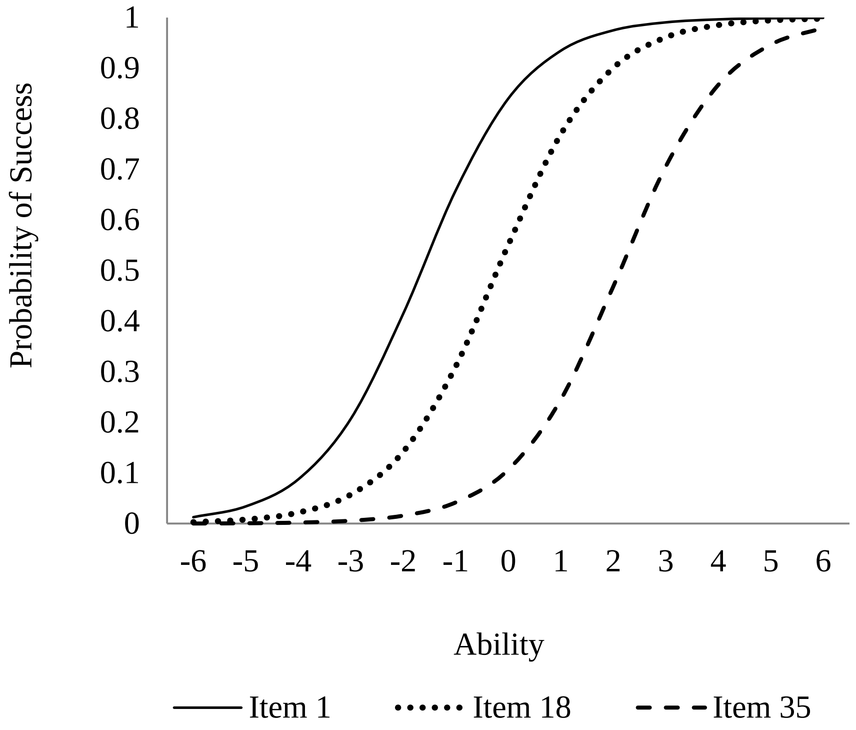


*Figure 3.* ICCs for three representative items from Sample 2 (combined group).

There were no further items showing misfit in either the male or female groups when calibrated separately. Two items showed significant DIF across sex, both of which were easier for males, with a difference of approximately .9 logits (Table 10). These items were different from those items showing DIF in Sample 1.

Table 10

*Item fit and DIF: Sample 2.*

| Item | Parameter Estimate | SE | Infit | t | Sex x item | SE | Wald t |
|---|---|---|---|---|---|---|---|
| 1 | -1.657 | .147 | 1.02 | 0.3 | -0.136 | .168 | -0.81 |
| 2 | -1.891 | .156 | 0.95 | -0.5 | 0.458 | .223 | 2.05* |
| 3 | -1.866 | .155 | 0.95 | -0.4 | -0.025 | .183 | -0.14 |
| 4 | -1.445 | .140 | 0.99 | -0.1 | 0.455 | .194 | 2.35* |
| 5 | -1.199 | .132 | 0.92 | -1.1 | -0.046 | .154 | -0.30 |
| 6 | -1.889 | .156 | 1.06 | 0.6 | 0.179 | .198 | 0.90 |
| 7 | -1.424 | .139 | 0.96 | -0.5 | 0.011 | .164 | 0.07 |
| 8 | -1.504 | .141 | 0.98 | -0.2 | -0.306 | .156 | -1.96 |
| 9 | -1.588 | .144 | 0.92 | -0.9 | 0.361 | .194 | 1.86 |
| 10 | -1.215 | .133 | 0.91 | -1.2 | 0.199 | .165 | 1.21 |
| 11 | -1.545 | .143 | 0.85 | -1.7 | -0.014 | .168 | -0.08 |
| 12 | -1.344 | .136 | 0.85 | -1.9 | 0.060 | .164 | 0.37 |
| 13 | | | | | | | |
| 14 | -1.287 | .135 | 1.02 | 0.3 | -0.207 | .151 | -1.37 |
| 15 | -0.892 | .125 | 1.01 | 0.2 | -0.275 | .138 | -1.99 |
| 16 | -0.876 | .125 | 0.94 | -0.9 | -0.144 | .141 | -1.02 |
| 17 | | | | | | | |
| 18 | -0.203 | .114 | 1.05 | 0.9 | -0.156 | .127 | -1.23 |
| 19 | -0.752 | .122 | 1.00 | 0.0 | -0.019 | .141 | -0.13 |
| 20 | | | | | | | |
| 21 | | | | | | | |
| 22 | 0.875 | .109 | 1.01 | 0.2 | 0.130 | .122 | 1.07 |
| 23 | 0.468 | .109 | 1.00 | 0.1 | -0.005 | .122 | -0.04 |
| 24 | 0.803 | .109 | 0.98 | -0.4 | 0.020 | .121 | 0.17 |
| 25 | 0.659 | .109 | 1.01 | 0.3 | 0.159 | .123 | 1.29 |
| 26 | 1.189 | .110 | 1.08 | 1.7 | 0.118 | .122 | 0.97 |
| 27 | 1.633 | .115 | 1.09 | 1.7 | -0.079 | .126 | -0.63 |
| 28 | | | | | | | |
| 29 | 2.335 | .128 | 1.12 | 1.6 | 0.096 | .136 | 0.71 |
| 30 | 1.541 | .113 | 1.01 | 0.2 | -0.176 | .125 | -1.41 |
| 31 | 1.580 | .114 | 1.00 | -0.0 | -0.116 | .126 | -0.92 |
| 32 | 1.619 | .115 | 1.02 | 0.3 | -0.056 | .126 | -0.44 |
| 33 | 1.864 | .119 | 1.07 | 1.1 | -0.243 | .131 | -1.85 |
| 34 | 1.907 | .119 | 1.01 | 0.2 | -0.248 | .132 | -1.88 |
| 35 | 2.127 | .124 | 0.91 | -1.4 | 0.086 | .132 | 0.65 |
| 36 | 4.065 | .205 | 1.03 | 0.2 | -0.081 | .220 | -0.37 |

**Sample 3.** One case showed misfit ($t$ = 13.62) and was excluded from analysis. All items showed acceptable fit to the Rasch model. The Rasch dimension explained 38% of the variance. The first eigenvalue from the PCA of Rasch residuals was 1.23, explaining 8.9% of the residual variance. This was below the cut-off value of 1.36, supporting unidimensionality of the measure. Although the variance explained by the Rasch dimension was lower than desirable, there was little evidence of the presence of a substantial second dimension. The female and male data also showed no evidence of multidimensionality (eigenvalues of 1.27 and 1.40 respectively, with cut-off values of 1.45 and 1.66). Figure 4 shows the ICCs for three representative items in Sample 3 for the combined group.

When the female and male data were considered separately, it was found that item 25 showed poor fit in the female group (infit = 1.13, $t$ = 2.2), while in the male group no items showed poor fit. There was no significant DIF across sex in the remaining items (Table 11).

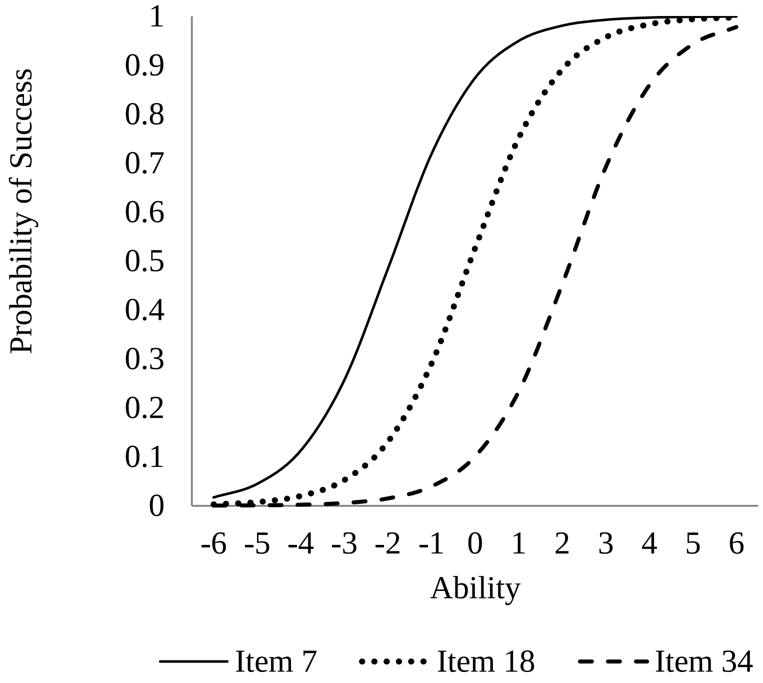


*Figure 4.* ICCs for three representative items from Sample 3 (combined group).

Table 11

*Item fit and DIF: Sample 3.*

| Item | Parameter | SE | Infit | t | Sex | SE | Wald t |
|---|---|---|---|---|---|---|---|
| | Estimate | | | | x item | | |
| 7 | -1.935 | .161 | 0.97 | -0.2 | -0.167 | .173 | -0.97 |
| 11 | -2.445 | .187 | 0.91 | -0.6 | 0.333 | .234 | 1.42 |
| 13 | -0.287 | .121 | 1.08 | 1.5 | -0.146 | .132 | -1.11 |
| 15 | -1.337 | .141 | 1.01 | 0.1 | -0.023 | .155 | -0.15 |
| 16 | -0.859 | .129 | 0.98 | -0.3 | -0.011 | .142 | -0.77 |
| 17 | -0.894 | .130 | 1.02 | 0.4 | -0.164 | .140 | -1.17 |
| 18 | -0.102 | .120 | .97 | -0.6 | -0.012 | .131 | -0.09 |
| 21 | 0.431 | .119 | .96 | -0.7 | 0.094 | .129 | 0.73 |
| 23 | 0.343 | .118 | .99 | -0.2 | 0.064 | .129 | 0.50 |
| 25 | 0.122 | .119 | 1.08 | 1.7 | | | |
| 26 | 0.892 | .121 | .98 | -0.4 | -0.087 | .130 | -0.67 |
| 27 | 1.172 | .124 | .90 | -1.8 | -0.243 | .134 | -1.81 |
| 30 | 0.861 | .121 | 1.1 | 1.8 | 0.101 | .130 | 0.78 |
| 32 | 1.846 | .137 | 1.08 | 1.1 | 0.279 | .142 | 1.96 |
| 34 | 2.194 | .146 | .87 | -1.6 | -0.019 | .152 | -0.13 |

## Discussion

Overall, the analyses presented largely support a unidimensional conceptualization of the APM, in the two short forms and the complete test, and in males and females. Furthermore, the analyses indicate that the APM is largely measurement invariant across sex and argue against any consistent sex differences in the item types identified to date.

Results regarding sex differences in items and item types displayed no clear pattern of association with item classifications from Carpenter et al. (1990), DeShon et al. (1995), or Dillon et al. (1981), and no clear pattern across samples. This is contrary to the findings from Mackintosh & Bennett (2005); however, it is consistent with findings from Colom and Abad (2007) and Vigneau and Bors (2008). There were significant sex differences in every item type in at least one of the samples and some individual items of all types demonstrated a significant sex difference. However, the lack of consistent results supports Vigneau and Bors’ (2008) contention that the use of these taxonomies to understand sex differences in APM items may not be a valuable line of enquiry.

It was hoped that analyses of the factor structure in males and females separately would provide new insight into both the structure of the test and sex differences in performance. However, the factor analytic results were largely in line with previous research that has looked at combined data only. This research has typically reported little support for models based on Dillon et al. (1981) and DeShon et al.’s (1995) distinctions (Abad et al., 2004; Alderton & Larson, 1990; Arthur et al., 1999; Vigneau & Bors, 2008) and found the best fitting model to be a single-factor model (Abad et al., 2004; Chiesi, Ciancaleoni, Galli, & Primi, 2012). The current study suggests these findings can be extended to males and females when considered separately. There was little support for Lim's (1994) finding of a different factor structure of the APM in males and females.

While there was some evidence of a multi-factor structure in the female group, this was only found in Sample 1. It is likely that this finding occurred due to the method of item selection used to create the Bors and Stokes (1998) short form of the test. They selected those items with the highest item-total correlation, but with low inter-item correlations, in order to remove any redundancy and to obtain a wide variety of different items. This is in contrast to the short form used in Sample 3, where items with the highest item-total and inter-item correlations were selected to create a more pure measure of the APM. Therefore, the test used in Sample 1 consists of a more dissimilar group of items than the original APM, and the test used in Sample 3 consists of a more similar group of items than the original APM. These two short forms had only 40-50% of items in common, and therefore were substantially different. Consequently, the method of item selection could explain why Sample 1 showed the greatest evidence of multidimensionality and Sample 3 showed the least.

The fact that the models based on item difficulty provided arguably the best fit is difficult to interpret, but is supported by previous findings indicating that a model based on item-skewness provided the best fit (Vigneau & Bors, 2008). This result may highlight some issues regarding the use of factor analytic methods to answer the question of dimensionality, or may be a real effect related to item position effects. There is some evidence that in addition to quantitative changes in the difficulty of items there may also be qualitative changes across the test (Vigneau & Bors, 2005), and this could also be responsible for the superior fit of the difficulty models. More research is required to disentangle the causes of any qualitative differences between the beginning and end of the test, and the influence of position effects. However, although the difficulty models fit the data well, the correlation between factors was high, particularly in Samples 2 and 3 (average correlation of .89), raising questions about the utility of considering two separate factors. Similar to the results providing some evidence of multidimensionality in the female data, the strongest evidence of multidimensionality concerning the difficulty models was found in Sample 1. Therefore, the factor analytic results indicated the presence of slight multidimensionality, but overall supported unidimensionality of the test, given the high correlations between factors in the well-fitting multi-factor models.

Interestingly, measurement invariance testing indicated that all forms were invariant across sex, with the exception of some variance in one item factor loading in each case. These loadings were higher in the male group than the female group. Therefore, there was some difference in the variance explained by the latent APM factor across sex. Additionally, latent mean difference testing supported the sex difference found in raw scores, and suggests that there is a sex difference in the latent construct measured by the APM. It is proposed that the sex difference in latent means can be interpreted one of three ways; either that males are simply better at reasoning, that visuospatial ability is involved, but was not able to be separated from reasoning ability in the current study and this is causing the difference, or that this finding is a result of the particular samples used (mainly Psychology students). Caution should be used when interpreting the results of this test until we can identify exactly why this difference occurs.

The Rasch analysis largely supported the factor analytic findings indicating unidimensionality, although again there was some evidence for slight multidimensionality. Item fit analysis indicated that the majority of items in all samples conformed to the measurement properties of the Rasch model. Between 8 and 14% of items displayed a slight deviation from the model, with 0-8% displaying high infit values, which is more problematic than low values in this case. A strict item fit criterion was used in this study, with the aim of detecting any slight departure from fit to the Rasch model. Therefore, the violations of fit were not large. The PCA of Rasch residuals also supported unidimensionality, finding no evidence of any substantial component remaining once the Rasch dimension was accounted for.

There was less evidence of significant DIF than found by Abad et al. (2004), with results more consistent with findings from Chiesi, Ciancaleoni, Galli, Morsanyi et al. (2012) and Chiesi, Ciancaleoni, Galli, and Primi (2012). Further, although there was some indication of significant DIF, the items identified as showing significant DIF in this study were neither the same as those showing DIF in Abad et al.'s study, nor the same as the items showing sex differences according to the chi-square tests in the current study. One interpretation of these results is that these different methods pick up on small variations between the sexes in slightly different ways. This interpretation is supported by the fact that, although there were some significant sex differences found, none of the effect sizes was particularly large and the findings were different across samples and methods. Furthermore, the fact that different items were found to show differences in the different samples can be interpreted in two ways; either the use of different forms may change the relationship between the items (as argued by Vigneau & Bors, 2005) or these analyses are picking up on slight variations within the sample, and there is no consistent difference.

Overall, it appears that, while the information processing taxonomies proposed by Carpenter et al., (1990) and DeShon et al., (1995), as well as the factors proposed by Dillon et al. (1981), may be useful for understanding the ways an individual can approach solution to APM items, they may not be helpful in distinguishing between items that do or do not involve visuospatial processes, or in understanding the mechanisms behind sex differences in performance. If the test is indeed multidimensional, then it may be that the item categorisations provided so far have not identified the factors causing multidimensionality. However, given the fit of the items to the Rasch model, if multidimensionality were present, it would likely not be a large effect. The present study supports previous literature suggesting that the test is unidimensional and has expanded this finding to suggest that this unidimensionality holds in both males and females, and is invariant across sex. This indicates that the test measures the same thing in both sexes, and that an overall score is sufficient in explaining performance on all forms of the test considered in the present study.

While the results of this study support the unidimensionality of the APM, this does not mean that the APM can definitively be said to not involve visuospatial ability. Unidimensionality does not, *per se*, lead to the conclusion that only one ability is involved in performance. Unidimensionality simply means that all items measure the same thing. Therefore, it is entirely possible that visuospatial ability is involved in the APM and that it is involved to a similar extent in most items. Research certainly suggests that visuospatial ability could play a role in APM performance; however, its overall importance may be questionable (Schweizer et al., 2007). Given this, in order to understand better how and to what extent APM performance may be related to visuospatial ability, more research involving administration of visuospatial ability tests in conjunction with the APM is needed. Investigating the dimensionality of the APM is a useful first step, but will not answer all questions regarding the relationship between the APM, visuospatial ability and sex differences.

# Chapter 4: Paper 2

## Preamble

The results of Paper 1 indicated that the influence of Gv may not be isolated in certain items, however this did not discount the possibility that visuospatial ability was involved in performance on the RPM tests as a whole, perhaps to a similar extent in all items. Unidimensionality does not signify that only a single construct is measured, simply that all items measure essentially the same thing.

Schweizer et al. (2007) performed an analysis investigating the contribution of Gv to performance on the APM beyond the contribution of Gf. Item parceling was used for this analysis in order to decrease the number of data points to be modeled. However, doing so results in the loss of item-specific information (Little et al., 2002), and may obscure any involvement of Gv processes. Therefore, we felt it was prudent to investigate this issue without the use of item parceling in order to either provide further support for Schweizer et al.'s findings or to provide evidence to the contrary.

One narrow Gv ability which shows particular promise as a candidate involved in RPM test performance is flexibility of closure, the ability to identify a stimulus embedded within a more complex figure (Schneider & McGrew, 2012). Flexibility of closure is conceivably related to element salience, a factor that has been linked to the difficulty of matrices items (Meo et al., 2007; M. J. Roberts et al., 2000). Although Schweizer et al. (2007) included three narrow abilities in their analysis (mental rotation, visualization and closure speed), flexibility of closure was not considered. Therefore, we felt that the additional consideration of this ability would add to the existing evidence.

Additionally, we thought it was important to assess the role of Gv in both the Advanced and Standard RPM. It has been argued that the Standard version of the RPM tests is more likely to be multidimensional (Colom & Abad, 2007), and therefore it may be more likely to involve Gv abilities than the Advanced versions. This could be because the relationships between the elements of the stimuli are more complex in the Advanced version, and therefore require higher-level reasoning ability. It is therefore useful to examine whether the Advanced and Standard RPM show a similar relationship to Gv abilities.

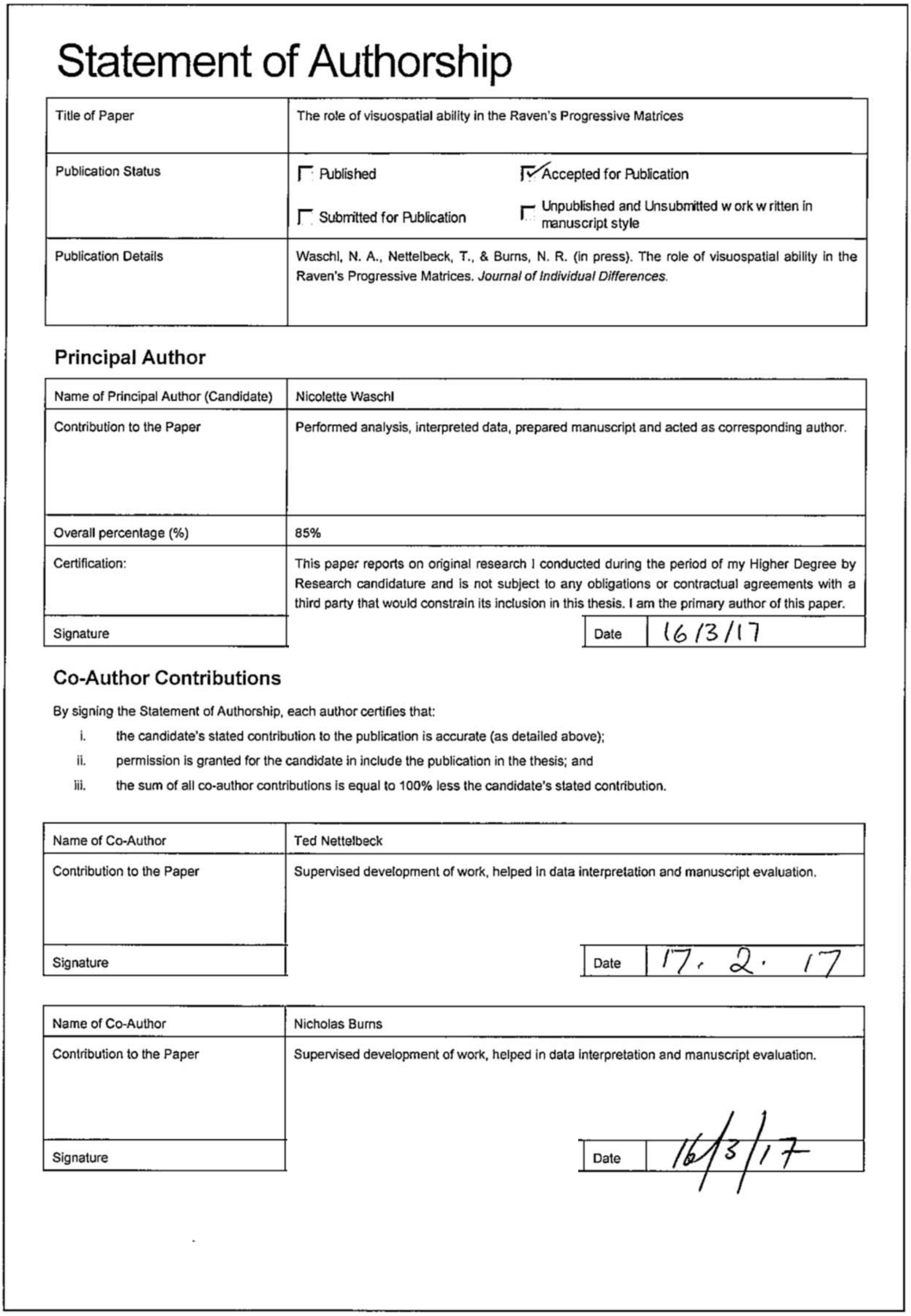
A final point that we felt warranted investigation was whether the relationship between Gf, Gv and the RPM was invariant across sex. Some research has indicated that RPM performance may be more closely related to Gv in females (Lim, 1994) and has identified different cortical regions recruited in APM performance across sex (Yang et al., 2014). Additionally, any sex differences in strategy use could influence the relationships between Gf, Gv and RPM. Although the literature regarding sex differences in strategy use is not entirely consistent, there is sufficient research reporting a significant sex difference (e.g. Geiser et al., 2006; Jansen-Osman, 2008) that we felt this issue warranted further investigation. Furthermore, if the structural parameters were found to be invariant, this would suggest a similar contribution of Gv to RPM performance across sex and would allow examination of whether a male advantage in Gv could account for the male advantage in latent RPM, as suggested by Colom, Escorial and Rebollo (2004).

# Paper 2

# The role of visuospatial ability in the Raven’s Progressive Matrices

Waschl, N. A., Nettelbeck, T., & Burns, N. R. (in press). The role of visuospatial ability in the Raven’s Progressive Matrices. *Journal of Individual Differences.*

*Note.* Supplementary materials can be found in Appendix B.



# Statement of Authorship

| | |
|---|---|
| Title of Paper | The role of visuospatial ability in the Raven's Progressive Matrices |
| Publication Status | □ Published<br>☑ Accepted for Publication<br>□ Submitted for Publication<br>□ Unpublished and Unsubmitted work written in manuscript style |
| Publication Details | Waschl, N. A., Nettelbeck, T., & Burns, N. R. (in press). The role of visuospatial ability in the Raven's Progressive Matrices. *Journal of Individual Differences.* |

## Principal Author

| | |
|---|---|
| Name of Principal Author (Candidate) | Nicolette Waschl |
| Contribution to the Paper | Performed analysis, interpreted data, prepared manuscript and acted as corresponding author. |
| Overall percentage (%) | 85% |
| Certification: | This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper. |
| Signature | Date 16/3/17 |

## Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

| | |
|---|---|
| Name of Co-Author | Ted Nettelbeck |
| Contribution to the Paper | Supervised development of work, helped in data interpretation and manuscript evaluation. |
| Signature | Date 17. 2. 17 |

| | |
|---|---|
| Name of Co-Author | Nicholas Burns |
| Contribution to the Paper | Supervised development of work, helped in data interpretation and manuscript evaluation. |
| Signature | Date 16/3/17 |

## Abstract

Debate surrounding the role of visuospatial ability in performance on the Raven's Progressive Matrices (RPM) has existed since their conception. This issue has yet to be adequately resolved, and may have implications regarding sex differences in scores. Therefore, this study aimed to examine the relationship between RPM performance, visuospatial ability and fluid ability, and any sex differences in these relationships. Data were obtained from three samples: two university samples completed the Advanced RPM and one population-based sample of men completed the Standard RPM. All samples additionally completed an alternative measure of fluid ability, and one or more measures of visuospatial ability. Structural equation modeling was used to examine the relationships between performance on the visuospatial and fluid ability tests and performance on the RPM. Visuospatial ability was found to significantly contribute to performance on the RPM, over and above fluid ability, supporting the contention that visuospatial ability is involved in RPM performance. No sex differences were found in this relationship, although sex differences in visuospatial ability may explain sex differences in RPM scores.

*Keywords:* Raven's Progressive Matrices; Fluid ability; Visuospatial ability; Sex differences

The Raven’s Progressive Matrices (RPM) have long been described as one of the best measures of general intelligence (e.g. Jensen, 1998), and are often used as such. However, debate exists surrounding what ability or abilities are involved in performance on these tests. Understanding this is integral to ensuring their construct validity. The RPM consist of figural matrices tasks, where each item is composed of a matrix of figural stimuli with an empty bottom right hand cell. The test taker must determine, out of eight response options, which option best fits the pattern and should fill the empty cell. To successfully solve these items, the individual is required to induce the rules governing the relationship between the stimuli and apply these rules to obtain the correct solution.

Three versions together form the RPM set of tests: the Coloured Progressive Matrices, for use with the intellectually disabled or young children; the Standard Progressive Matrices (SPM), for use with the general population; and the Advanced Progressive Matrices (APM), for use with university students, or those of above average intellectual ability. The main difference between these three versions is their difficulty and the complexity of the relationships to be induced. The present study is concerned with the SPM and APM, while the term RPM is used to refer to the set of tests.

Although the RPM were originally conceptualised as a measure of general intelligence, there are more precise ways of conceptualising the specific ability or abilities they measure. The Cattell-Horn-Carroll (CHC) model (Schneider & McGrew, 2012) is a three-stratum model of intelligence consisting of narrow abilities, subsumed by broad abilities, and a general intelligence factor. This theory was based on Cattell and Horn’s broad factors (Horn, 1965, 1994), and integrates the factor-analytic work of Carroll (1993). There is excellent empirical and theoretical support for CHC theory

(Ortiz, 2015) and the RPM will be considered under this taxonomy. The RPM are considered a measure of the CHC broad factor fluid ability (Gf), the ability to solve unfamiliar problems without relying on previous knowledge. More specifically, they measure inductive reasoning, the ability to determine the underlying principles or rules of the relationship between unfamiliar stimuli (Schneider & McGrew, 2012).

Although the RPM are excellent measures of Gf, some have argued that successful performance also involves visuospatial ability (e.g. DeShon, Chan & Weissbein, 1995). Visuospatial ability (Gv) is the ability to solve problems using visual images, and involves the generation, retention, retrieval and transformation of information presented in visuospatial form (Schneider & McGrew, 2012). Gv is not simply the ability to encode visuospatial stimuli, but involves the use of higher-level processes required for tasks such as object recognition and mental manipulation. Carroll (1993) presented evidence for at least five narrow Gv abilities, including visualization, speeded rotation, closure speed, flexibility of closure and perceptual speed.

While the CHC model differentiates Gf and Gv at the same broad level, alternative models of intelligence provide additional insights into the relationship between the use of visual stimuli and reasoning abilities. The Berlin model of Intelligence Structure (BIS; Jäger, 1982; Süβ  & Beauducel, 2005) is a faceted model of intelligence incorporating both content and process dimensions of cognitive abilities, where the process represents the cognitive processes involved (e.g. reasoning, memory) and the content represents the stimulus used to assess the cognitive process (figural, verbal, and numeric). Under the BIS, the RPM measure figural-spatial reasoning (Süß & Beauducel, 2015), indicating the importance of the visual stimuli. Interestingly, Carroll’s (1993) analysis indicated that the distinction between the three lower-order Gf factors of induction, general sequential reasoning and quantitative reasoning correspond to the content factors of figural, verbal and quantitative, and that the narrow Gv ability of visualization is closely associated with Gf.

Recently, some further support has been found for the division of reasoning abilities into content-specific sub-factors using the hierarchies of factor solutions approach to examine the structure of cognitive abilities (Lang, Kersting & Beauducel, 2016). These analyses also indicated that Gf, or reasoning, factors appear at higher levels of the cognitive ability hierarchy than do factors related to perceptual abilities. Perceptual abilities tended to appear at the fifth level of the hierarchy, and were most highly correlated with reasoning and memory abilities. This suggests that visuospatial abilities may be more strongly related to Gf or reasoning than other abilities.

Several experimental and psychometric studies have provided corroborating evidence that Gv is involved in RPM performance, for both the APM and SPM. DeShon et al. (1995) reported that concurrent verbalization of the solution strategy used to solve APM items resulted in decrements in performance on certain items, suggesting that verbalization interfered with the visual processes required for item solution. Stephenson and Halpern (2013) investigated the effect of different types of working memory training on the APM, and found that all types of training including a visuospatial component improved performance above that of a control group, while training that did not include a visuospatial component did not. Regarding the SPM, Lynn, Allik, and Irwing (2004) identified three latent factors underlying test performance - gestalt continuation, verbal-analytic and visuospatial – again suggesting a role for Gv in SPM performance. It is also not uncommon that factor analyses of large test batteries find the RPM to load on a combined Gf/Gv factor (e.g. Crawford, 1991; Roberts & Stankov, 1999).

Research into the cognitive processes involved in completing matrices tasks also provides evidence that Gv may be involved. Perceptual organization of the items explains a large part of the variance in item complexity (Primi, 2001) and is associated with decreases in the latent Gf saturation of items (Arendasy & Sommer, 2012). Given that a key component of Gv involves the transformation of visuospatial information, the complexity of the perceptual organization of items could implicate Gv. Additionally, research using computational modeling to simulate the solution processes involved in RPM performance has found that visual representational strategies can be successfully used (Kunda, McGreggor & Goel, 2013).

One study that has directly investigated the variance in RPM performance explained by Gv was Schweizer, Goldhammer, Rauch, and Moosbrugger (2007). These authors used structural equation modeling to explore the relationship of three measures of Gv and one of inductive reasoning to APM performance. Whether the Gv measures were considered separately, or a latent Gv factor used, Gv was not significantly predictive of APM performance after accounting for inductive reasoning. However, Schweizer et al.'s analyses were carried out using item parceling. Three parcels for each scale were calculated systematically: the first parcel consisted of every third item (first, fourth, seventh etc.), the second parcel consisted of each following item (second, fifth, eighth etc.), and the third again consisted of each following item (third, sixth, ninth etc.). Each latent variable was then modeled based on the total scores for each of the three relevant item parcels. While there are arguments both for and against item parceling, it can be problematic when there is evidence of test multidimensionality (Little, Cunningham, Shahar & Widaman, 2002). Given the uncertainty surrounding the dimensionality of the RPM (e.g. Lynn, Allik & Irwing, 2004; Vigneau & Bors, 2005), this is important. Further, while Schweizer et al.'s results provide evidence against the involvement of Gv in APM performance, a similar study has yet to be conducted with SPM data. Therefore, the current study sought to add to Schweizer et al.'s work through replicating the APM analysis, but modeling individual items rather than item parcels, as well as extending this analysis to the SPM.

If the RPM involve Gv, there may be a specific narrow Gv ability implicated. Three narrow abilities present likely candidates: visualization, rotation, and flexibility of closure. Visualization concerns the ability to perceive and mentally reproduce how patterns might look when transformed (Carroll, 1993). Included under visualization are tests of mental rotation, which tend to show very large sex differences (Linn & Petersen, 1985; Voyer, Voyer & Bryden, 1995) and because of this unique characteristic, are often viewed as slightly different and afforded their own category in analysis. Given that the solution of RPM items can rely partially on transforming and rotating certain elements of the stimulus, visualization and rotation present promising candidates. Flexibility of closure concerns the ability to identify a stimulus embedded within a complex figure (Schneider & McGrew, 2012). There is evidence that element salience, or the ease of identification of specific elements, is an important difficulty factor in RPM items (Meo, Roberts & Marucci, 2007; Roberts, Welfare, Livermore & Theadom, 2000), indicating a role for flexibility of closure in performance. In their study, Schweizer et al. (2007) used measures of visualization, rotation and closure speed. Closure speed is different from flexibility of closure, involving identification of a familiar and meaningful object from an incomplete stimulus (Schneider & McGrew, 2012). Arguably, the inclusion of a measure of flexibility of closure may be more relevant to the type of abilities hypothesized to be related to RPM performance. This ability is therefore included in the present study.

Finally, one pertinent concern regarding the involvement of Gv in RPM performance is that males tend to perform better on certain tests of Gv (Linn & Petersen, 1985; Voyer et al., 1995). This could have important implications if the RPM involve Gv and interpretation of scores does not take this into account. For example, sex differences in RPM scores are sometimes claimed to represent sex differences in general intelligence (e.g. Lynn & Irwing, 2004). However, research indicates sex differences in Gv may be responsible for sex differences on the APM (Colom, Escorial, & Rebollo, 2004), which could be problematic for this interpretation. Therefore, it is important to consider how sex is implicated in the relationship between RPM performance, Gf and Gv.

This study will use three data sets to investigate the three issues discussed here: two data sets concern the APM, and one the SPM. First, the utility of Gv in explaining variance in RPM scores (both APM and SPM) will be investigated. For the APM data this represents a replication of Schweizer et al.'s (2007) work but conducted with different measures of the relevant constructs, and conceptualized and modeled slightly differently. For the SPM data this closely follows Schweizer et al., but using the SPM instead of the APM. Secondly, if Gv does contribute to RPM performance, the role of narrow Gv abilities will be considered. Finally, this study will investigate sex differences in the relationship between Gv and APM performance.

# Method

## Participants

**APM Sample 1.** Participants were *N* = 353 undergraduate psychology students from the University of Sydney, Australia, recruited over two studies (Jackson, Kleitman, Stankov & Howie, 2016). Each participant completed the three relevant cognitive ability measures as part of a larger battery of tests. The sample consisted of 112 males and 241 females, aged from 17 to 54 years (*M* = 20.1, *SD* = 3.20).

**APM Sample 2.** Participants were *N* = 236 undergraduate psychology students from McGill University, Canada. The sample consisted of 65 males and 171 females (four participants with no gender data were excluded). Age ranged from 17 to 32 years (*M* = 20.8, *SD* = 1.89).

**SPM Sample.** Participants were 287 individuals from a population representative sample of men in Adelaide, Australia, who completed the five cognitive ability measures reported here as part of a larger battery of tests. The age of participants ranged from 37 to 83 years (*M* = 60.6, *SD* = 11.37). For further details of the original study see Kelly, Burns, Bradman, Wittert and Daniel (2012).

## Measures

**APM Sample 1.** Participants completed a 15-item short form of the APM (Raven, Court & Raven, 1998). This short form consisted of items 7, 11, 13, 15, 16, 17, 18, 21, 23, 25, 26, 27, 30, 32 and 34 from Set II of the full form APM, and was based on extensive pilot testing which demonstrated appropriateness of the selected items for the target population.

To measure Gf, the Esoteric Analogies (Stankov, 1997) test was used. This test consisted of 20 items for which participants had to choose one word to complete the verbal analogy. Items take the format X is to Y as Z is to: A, B, C or D? (e.g. Fire is to Ice as Hot is to: Warm, Cold, Orange, Blue?).

To measure Gv, a 15-item short form of the Purdue Spatial Visualization Test of Rotations: Revised (PSVT:R; Yoon, 2011) was used. Each item consists of an example of a figure and its corresponding form when rotated at a given angle. A stimulus figure is then presented, and participants must choose out of five response options which target figure demonstrates the same degree of rotation of the stimulus figure as that shown in the example pair. Questions are presented in the format X is rotated to Y as Z is rotated to A, B, C, D or E? This test was designed to measure mental rotation.

Measures were completed online in participants' own time or as part of their tutorial program. There was no indication of any differences in the construct measured depending on whether the measure was completed online or in the tutorial. All measures were administered untimed, and participants were not permitted to revise their responses once selected.

**APM Sample 2.** Participants completed a 12-item short form of the APM (Bors & Stokes, 1998), with a 15-minute time limit. All items were from Set II of the full form. To measure Gf, the Comprehensive Ability Battery – Inductive Reasoning test (CAB-I; Hakstian & Cattell, 1975) was used. Each item in this test consists of five letter sets. The task is to determine the pattern across the letter sets, and to identify which set does not follow the same pattern as the others. Participants were allowed six minutes to complete the 12 items.

To measure Gv, the Mental Rotations Test (MRT; Vandenberg & Kuse, 1978) and the Differential Aptitude Tests: Space Relations (Paper Folding; Bennett, Seashore & Wesman, 1989) tests were used. The MRT measures mental rotation. Each item consists of five three-dimensional figures: one target figure and four similar figures. Participants must identify which two of these four similar figures are the same as the target figure, but rotated in space. This test was divided into two sections consisting of 10 questions each and participants were allowed three minutes to complete each section. The Paper Folding test measures visualization. Each item consists of a target pattern, and four drawings of three-dimensional figures. Participants are required to determine which of these three-dimensional figures could be made from the target pattern. A 12-minute time limit was used for this measure.

All measures were completed in paper-and-pencil form in a group setting and participants were permitted to revise their responses at any point within the time limit.

**SPM Sample.** Participants completed the SPM (Raven, Raven & Court, 1998). This version of the RPM consists of 60 items, divided into five sets. To measure Gf, the Comprehensive Ability Battery – Inductive Reasoning test (Hakstian & Cattell, 1975) was used, and to measure Gv, the Comprehensive Ability Battery – Flexibility of Closure test (CAB-Cf; Hakstian & Cattell, 1975), Mental Rotations Test (MRT; Vandenberg & Kuse, 1978) and the Differential Aptitude Tests: Space Relations (Paper Folding; Bennett, Seashore & Wesman, 1989) tests were used. The CAB-Cf measures flexibility of closure. Each item requires participants to identify a known simple figure disguised within a complex figure. MRT and Paper Folding measures were as described in Section 2.2.2. All measures were administered timed; 20 minutes was given to complete the RPM, six minutes for the CAB-I and CAB-Cf (12 items), five minutes for the Mental Rotations Test (20 items) and 12 minutes for the Paper Folding test (30 items). All measures were completed in paper-and-pencil form, in a group setting, and participants were permitted to revise their responses at any point within the time limit.

## Design

As Schweizer et al. (2007) argued, to effectively investigate the contribution of Gv to RPM performance over Gf, an appropriate alternative test of Gf that does not involve Gv to any substantial extent must be used. The Esoteric Analogies test is a factorially complex measure, typically found to load both Gf and crystallised ability (Gc; the ability to use experience and learned knowledge; Schneider & McGrew, 2012**)**, with the relative importance of these broad abilities changing depending on the test battery used in analysis. Roberts and Stankov (1999) found that when a Gf/Gv factor was identified in their test battery, Esoteric Analogies loaded only onto Gc. Given that other studies have found combined loadings of the Esoteric Analogies on both Gf and Gc (Kleitman & Stankov, 2007; MacCann, 2010), this indicates that the measure does not involve Gv. The CAB-I is a measure of inductive reasoning that does not consist of figural stimuli. Hence, it is a test highly related to the RPM, but unlikely to involve Gv.

## Analysis

Structural equation modeling was used to examine the relationships between latent reasoning and visuospatial factors derived from the various cognitive ability tests, and a latent APM or SPM factor. All structural equation modeling was performed in Mplus 7 (Muthén & Muthén, 1998-2012). Models were initially computed using individual items, given concerns regarding item parcelling. However, in order to compare results to those of Schweizer et al. (2007) and consider the influence of item parceling on results, analyses using item parcels were also completed. These were conducted as described by Schweizer et al., with three item parcels used for each factor. The weighted least squares mean and variance adjusted estimator (WLSMV), which has been shown to perform well with dichotomous data (Muthén, du Toit, & Spisic, 1997), was used for the individual item analysis, and the maximum likelihood (ML) estimator for the item parcel analysis.

A two-step modeling procedure (Anderson & Gerbing, 1988) was followed, whereby the measurement model is evaluated first to establish good fit, and following this the structural parameters are added. Model fit was evaluated with reference to the chi-square value, RMSEA, CFI and SRMR. Guidelines for interpreting these indices under the WLSMV estimator recommend the following cut-off values for acceptable fit: RMSEA ≤ .05; CFI ≥ .96 and SRMR ≤ .07 (Yu, 2002). For the ML estimator, guidelines recommend the following cut-off values: normed chi-square (i.e. $\chi^2$/df) < 2 (Kline, 1998), RMSEA < .05 (Browne & Cudeck, 1993), CFI ≥ .95 and SRMR < .08 (Hu & Bentler, 1999).

**Incremental validity models.** The current study departs from Schweizer et al.'s (2007) in the manner of modeling the relationship between Gf, Gv and RPM. The current study was specifically interested in investigating the extent to which Gv abilities may explain the variance in RPM not already accounted for by Gf. Thus, the modeling of the relationships between Gf, Gv and latent RPM differs conceptually from the models used in Schweizer et al. in order to explicitly show this, but is functionally equivalent. In the present study, a latent RPM factor was modeled and regressed on the Gf factor. A "residual latent RPM" factor was then created to represent the variance in latent RPM not explained by Gf. This residual factor was regressed on the Gv factors.

These models can be understood as analogous to an incremental validity analysis with manifest variables[1]. Modeling in this way allows more explicit consideration of the extent to which Gv accounts for the residual RPM variance, specifically. This is important, because the RPM are already a well-established measure of Gf, and thus the extent to which Gf predicts latent RPM is of less concern than the extent to which Gv explains the remaining variance.

**Multiple group confirmatory factor analysis.** Multiple group confirmatory factor analysis was used to examine invariance across sex in the strength of the relationships between Gf, Gv and RPM. If the structural parameters were found to be invariant across sex, this would indicate that sex does not moderate the relationship between Gv and RPM.

Testing for measurement invariance in structural parameters involved first establishing that the measurement model was invariant across sex. With categorical data, this involves the comparison of a less restrictive and more restrictive model; tested by constraining equal the item thresholds and factor loadings. If there is no significant difference in model fit according to the chi-square statistics, this indicates measurement invariance. The Mplus DIFFTEST function (Asparouhov & Muthén, 2006) was used to test for these differences. This function follows a two-step process whereby, first, the less restrictive model is estimated and the derivatives needed for the chi-square difference test are saved. The more restrictive model is then estimated and the chi-

[1] Incremental validity analyses of manifest variables can be found in the supplementary material.

square difference test calculated using the derivatives from both models (Muthén & Muthén, 1998-2012).

**Multiple indicators multiple causes modeling.** In order to investigate whether a sex difference in latent Gv might be responsible for the sex difference in APM, Multiple Indicators Multiple Causes (MIMIC) modeling was used. This approach to modeling tests the effects of a covariate, in this case sex, on the latent variables. In order to examine sex as an antecedent to the Gv-RPM relationship, a model in which RPM and Gv were both regressed on sex was compared with a model in which only RPM was regressed on sex. If the sex difference in Gv accounted for the sex difference in RPM, the parameter from sex to RPM would become non-significant once accounting for the parameter from sex to Gv.

## Results

### Descriptive Statistics

**APM Sample 1.** Table 1 shows the means and standard deviations on all measures for males and females. All measures demonstrated a significant sex difference. As expected, the largest difference was for Rotations.

Correlations between measures are presented in Table 2 for males and females separately. Correlations between the APM and Rotations were of a similar magnitude for males and females; however, correlations between the other tests were generally lower for females. Specifically, Analogies was not as strongly correlated with APM in females. However, the bootstrapped 95% confidence intervals (calculated using the bias-corrected and accelerated [BCa] method) of the correlations showed that all confidence intervals overlapped, indicating little difference between correlations across groups. A formal test of the difference between each pair of correlations indicated the only correlation to differ significantly between groups was between APM and Analogies (see Table 2).

**APM Sample 2.** Table 1 shows the means and standard deviations on all measures for males and females. There was a significant difference between the scores of males and females on the APM and MRT.

The correlations between measures are presented in Table 3. Some correlations in the male group were surprisingly small and non-significant, possibly due to the small number of males in this sample. Although many of the male correlations were quite low, all 95% confidence intervals overlapped with those of the females. Formal testing of the difference between pairs of correlations indicated no correlation differed significantly between groups. In both groups, CAB-I showed the strongest correlation with APM, as expected.

**SPM Sample.** Table 1 shows the means, standard deviations and ranges for all measures. Score ranges on the MRT and Paper Folding (PF) tests show participants tended to score low on these measures, most likely because the sample was population representative and were, on average, older (*M* = 60.6 years). Table 3 shows the correlations between measures in the SPM sample. The CAB-I showed the highest correlation with the SPM, as expected.

Table 1

*Descriptive Statistics*

| | Male | | | | Female | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | M | SD | *N* | Range | M | SD | *N* | Range | *α* | *d* | t | *p* |
| Sample 1 | | | 112 | | | | 241 | | | | | |
| APM | 8.97 | 3.73 | | 1 – 15 | 8.04 | 3.32 | | 0 – 15 | .80 | .27 | 2.36 | .019 |
| Rotations | 9.59 | 3.38 | | 2 – 15 | 7.98 | 3.28 | | 0 – 14 | .75 | .49 | 4.25 | < .001 |
| Analogies | 7.97 | 2.7 | | 2 – 13 | 7.11 | 2.67 | | 0 – 13 | .65 | .32 | 2.83 | .005 |
| Sample 2 | | | 65 | | | | 171 | | | | | |
| APM | 8.91 | 2.24 | | 1 – 12 | 8.03 | 2.41 | | 2 – 12 | .68 | .37 | 2.55 | .012 |
| CAB-I | 9.58 | 1.80 | | 4 – 12 | 9.07 | 1.98 | | 2 – 12 | .58 | .26 | 1.83 | .069 |
| MRT | 10.37 | 4.86 | | 1 – 20 | 7.05 | 4 | | 0 – 19 | .86 | .78 | 5.35 | < .001 |
| PF | 21.05 | 5.13 | | 10 – 30 | 19.60 | 5.08 | | 4 – 30 | .86 | .29 | 1.96 | .051 |
| Sample 3 | | | 287 | | | | | | | | | |
| SPM | 39.0 | 8.90 | | 10 – 57 | | | | | .92 | | | |
| CAB-Cf | 6.96 | 2.96 | | 0 – 12 | | | | | .81 | | | |
| MRT | 3.77 | 2.37 | | 0 – 11 | | | | | .60 | | | |
| PF | 7.84 | 3.36 | | 0 – 18 | | | | | .74 | | | |
| CAB-I | 3.70 | 2.02 | | 0 – 10 | | | | | .62 | | | |

*Note.* CAB-I = Comprehensive Ability Battery – Inductive Reasoning; CAB-Cf = Comprehensive Ability Battery – Flexibility of Closure; MRT = Mental Rotations Test; PF = Space Relations: Paper Folding

Table 2

*Correlations: APM Sample 1*

| | APM | Analogies | Rotations |
|---|---|---|---|
| APM | | .47[a] | .63 |
| | | *(.36 - .56)* | *(.54 - .70)* |
| Analogies | .64[a] | | .40 |
| | *(.53 - .73)* | | *(.29 - .51)* |
| Rotations | .63 | .55 | |
| | *(.47 - .71)* | *(.38 - .65)* | |

*Note.* All correlations are significant at $p < .001$; bootstrapped 95% CIs presented in italicized brackets below correlation; Female correlations are presented above the diagonal, male correlations are presented below the diagonal; [a] indicates correlations statistically significantly differ ($z = 2.15$, $p = .03$).

Table 3

*Correlations: APM Sample 2 and SPM Sample*

| | APM/SPM | CAB-I | MRT | PF |
|---|---|---|---|---|
| APM Sample 2 | | | | |
| APM | | .48*** | .41*** | .47*** |
| | | *(.34 - .59)* | *(.29 - .52)* | *(.34 - .58)* |
| CAB-I | .39** | | .31*** | .47*** |
| | *(.18 - .57)* | | *(.18 - .41)* | *(.34 - .57)* |
| MRT | .29* | .18 | | .57*** |
| | *(.03 - .52)* | *(-.03 - .39)* | | *(.43 - .66)* |
| PF | .25 | .23 | .60*** | |
| | *(.02 - .44)* | *(-.03 - .47)* | *(.44 - .73)* | |
| SPM Sample | | | | |
| CAB-I | .64*** | | | |
| | *(.57 - .69)* | | | |
| MRT | .46*** | .37*** | | |
| | *(.36 - .54)* | *(.26 - .47)* | | |
| PF | .41*** | .36*** | .39*** | |
| | *(.31 - .50)* | *(.25 - .46)* | *(.28 - .50)* | |
| CAB-Cf | .58*** | .47*** | .47*** | .36*** |
| | *(.50 - .63)* | *(.37 - .55)* | *(.38 - .55)* | *(.25 - .47)* |

*Note.* ** $p$ < .01, *** $p$ < .001; bootstrapped 95% CIs presented in italicized brackets below correlation; Female correlations are presented above the diagonal, male correlations are presented below the diagonal; CAB-I = Comprehensive Ability Battery – Inductive Reasoning; CAB-Cf = Comprehensive Ability Battery – Flexibility of Closure; MRT = Mental Rotations Test; PF = Space Relations: Paper Folding

## Structural Equation Models

**APM Sample 1.** The measurement portion of this model consisted of three factors: rotation, reasoning and latent APM, with the items from each test loading onto their respective factors. All items showed significant loadings on their factors, with the exception of analogies item 12 and rotations item 15, which were subsequently excluded. The overall measurement model showed good fit to the data ($\chi^2_{(816)}$ = 912, *p* = .010 [normed chi-square = 1.12]; RMSEA = .02, CFI = .98, SRMR = .08).

The structural parameters were then added to this model to examine: (i) how well reasoning accounted for the latent APM factor; and (ii) the proportion of the remaining variance (i.e. the residual variance of latent APM not accounted for by reasoning), accounted for by rotation. This involved first regressing the latent APM factor on the Gf factor to determine how well Gf accounted for the variance in latent APM. Following this, the residual variance of latent APM remaining after accounting for the variance explained by the Gf factor was modeled and regressed on the Gv factor to investigate how well Gv accounted for this remaining variance.

Figure 1 shows that rotation accounts well for that portion of variance that reasoning does not account for in the latent APM factor, $R^2$ = .52, $p < .001$. Rotation explains over half of the variance remaining in latent APM once the variance accounted for by reasoning has been removed. A functionally similar model where the latent APM factor was regressed on both reasoning and rotation simultaneously showed that rotation may be more important in predicting latent APM than reasoning (γ = .54 and .39, respectively). This model also explained more variance in latent APM ($R^2$ = .74) than did the model including reasoning only ($R^2$ = .60).

Results from the equivalent models using item parcels were similar (see Figure 1). All item parcels showed acceptable loadings on their respective factors and the overall measurement model demonstrated good fit to the data ($\chi^2_{(24)}$ = 44.38, *p* = .007 [normed chi-square = 1.85]; RMSEA = .05, CFI = .98, SRMR = .03). The parameter estimates obtained were similar to those found in the equivalent models using individual items. As with the individual item model, the functionally similar model indicated that Gv may be more important in predicting latent APM than reasoning.

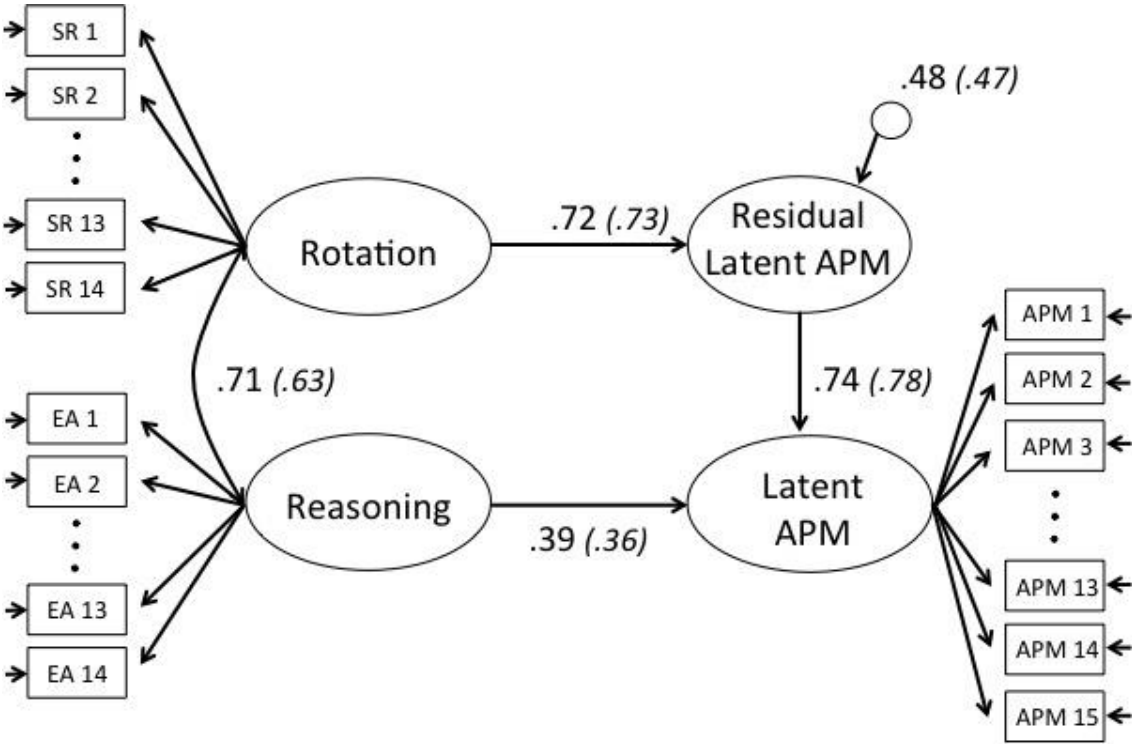


*Figure 1.* APM Sample 1 Model: Structural equation model showing standardized coefficients for reasoning and rotation to latent APM and residual latent APM. Values in parentheses indicate parameters when item parceling was used.

**APM Sample 2.** Several MRT, Paper Folding and CAB-I items were not included in the APM Sample 2 analysis due to several pairs of items being statistically indistinguishable. All other items demonstrated acceptable loadings on their respective factors.

Four models were tested in this sample; one in which a latent broad Gv factor predicted the residual of the latent APM variable; one in which both Gv abilities independently but simultaneously predicted the residual of the APM latent variable; and two in which each individual Gv ability independently predicted the residual of the APM latent variable. The first model was a test of the extent to which broad Gv influences APM, while the remainder allowed consideration of which narrow ability (rotation or visualization) might demonstrate the strongest prediction of APM. Because of the problem of correlated predictors, Models 3-4 were included to examine narrow Gv abilities separately.

The measurement model for Model 1 consisted of latent variables for APM, inductive reasoning (defined by CAB-I items), rotation (MRT items) and visualization (Paper Folding items). The mental rotation and visualization factors then defined a broad Gv factor. Model 2 consisted of latent APM and both latent Gv abilities, without the broad Gv factor. Models 3 – 4 consisted of latent APM and rotation (Model 3) or visualization (Model 4). Table 4 contains the fit indices for these models. CFI values were lower than the recommended ≥ .96, and SRMR values were higher than the recommended ≤ .07, but given the acceptable chi-square and RMSEA values, and CFI values close to .96 in most cases, examination of structural parameters proceeded.

Examination of Model 1 showed that the latent Gv factor significantly predicted residual latent APM ($R^2 = .34$, $p = .006$; Figure 2). In Model 2 (Figure 3) only rotation significantly predicted residual latent APM, but together rotation and visualization predicted nearly one quarter of variance in residual latent APM ($R^2 = .22$, $p = .009$). Inspection of Models 3-4 showed that neither mental rotation ($R^2 = .16$, $p = .029$) nor visualization ($R^2 = .16$, $p = .054$) were able to explain residual latent APM as well as the broad Gv factor.

Table 4

*Fit Statistics for Measurement Models: APM Sample 2*

| Model | $\chi^2$ *(df)* | Normed $\chi^2$ | RMSEA | CFI | SRMR |
|---|---|---|---|---|---|
| Individual Items | | | | | |
| Model 1: Broad Gv | 1949 *(1705)* | 1.14 | .03 | .94 | .13 |
| Model 2: Both narrow Gv abilities | 1948 *(1704)* | 1.14 | .03 | .94 | .13 |
| Model 3: Mental Rotation | 796 *(557)* | 1.43 | .04 | .90 | .13 |
| Model 4: Visualization | 1056 *(899)* | 1.17 | .03 | .93 | .13 |
| Item Parcels | | | | | |
| Model 1: Broad Gv | 60.50 *(49)* | 1.23 | .03 | .99 | .03 |
| Model 2: Both narrow Gv abilities | 58.25 *(48)* | 1.21 | .03 | .99 | .03 |
| Model 3: Mental Rotation | 25.62 *(24)* | 1.07 | .02 | 1.00 | .03 |
| Model 4: Visualization | 18.23 *(24)* | 0.76 | < .01 | 1.00 | .02 |

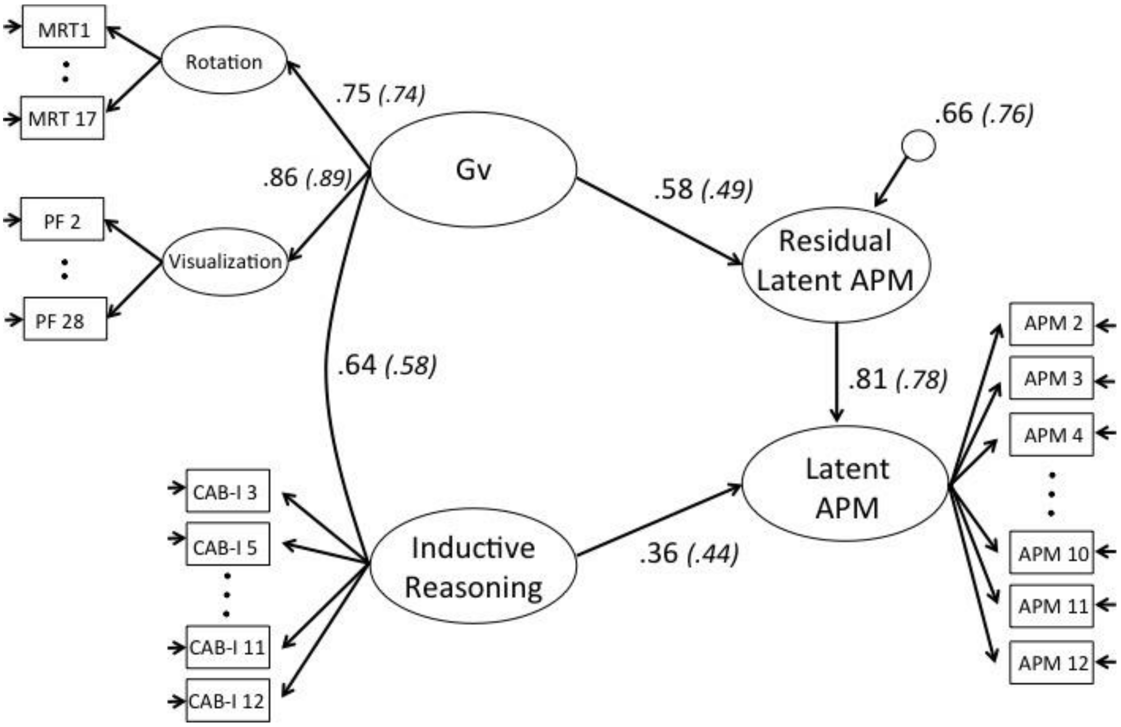


*Figure 2.* APM Sample 2 Model 1: Structural equation model showing standardized coefficients for latent Gv and inductive reasoning to latent APM and residual latent APM. Values in parentheses indicate parameters when item parceling was used.

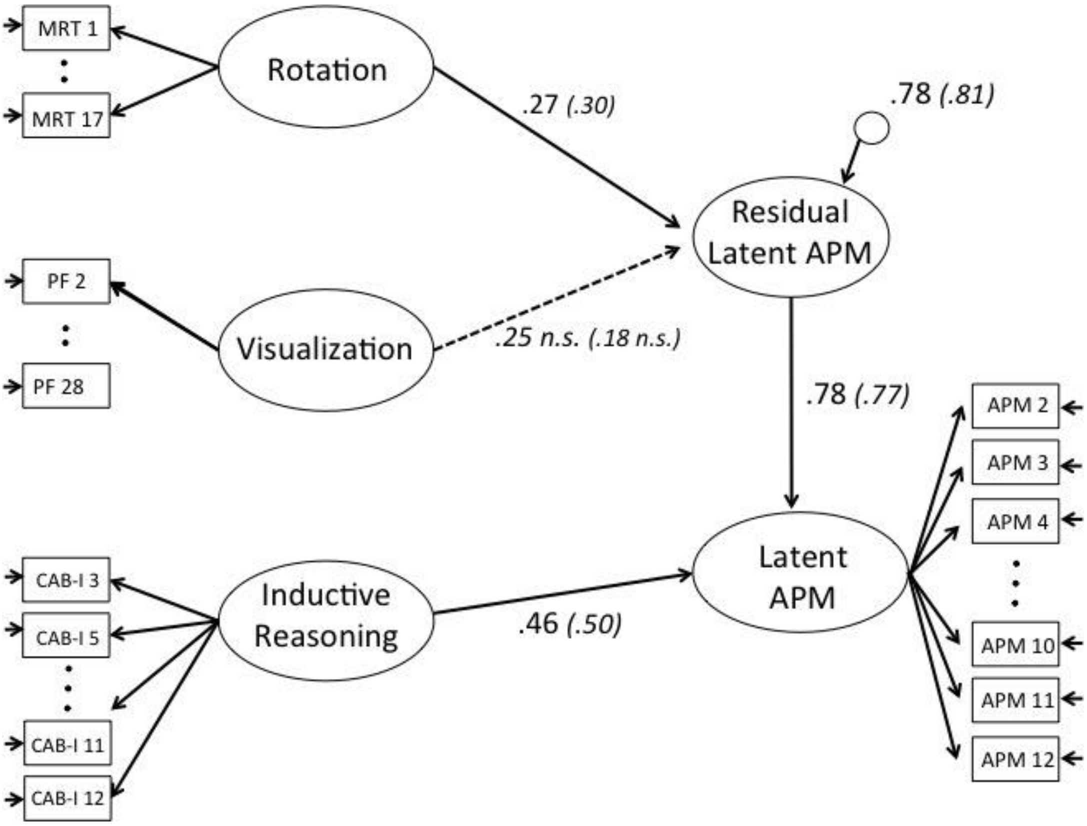


*Figure 3.* APM Sample 2 Model 2: Structural equation model showing standardized coefficients for ability measures to latent APM and residual latent APM. Correlations between factors were modeled but not depicted here for diagram clarity. Values in parentheses indicate parameters when item parceling was used.

In Models 1 and 2, broad Gv and the two individual Gv abilities predicted less well the variance in residual latent APM not accounted for by inductive reasoning than was found in APM Sample 1. However, the Gv abilities were still able to predict up to one third of the residual APM factor. The broad Gv factor accounted for this variance somewhat better than the combined individual Gv abilities.  A functionally similar model where latent APM was regressed on both inductive reasoning and latent broad Gv simultaneously indicated that broad Gv may be more important in predicting latent APM than reasoning ($\gamma$ = .47 and .36, respectively); however the other three models showed the opposite result: Model 2 ($\gamma$ = .19 [Visualization], $\gamma$ = .21 [Mental Rotation] and $\gamma$ = .46 [Inductive Reasoning]); Model 3  ($\gamma$ = .30 [Mental Rotation] and $\gamma$ = .53 [Inductive Reasoning]); and Model 4 ($\gamma$ = .31 [Visualization] and $\gamma$ = .49 [Inductive Reasoning]). The models including broad Gv and both narrow Gv abilities explained more variance in latent APM ($R^2$ = .57 and .53, respectively) than did the model including inductive reasoning only ($R^2$ = .45), while the difference was smaller in those models containing only a single Gv ability (Model 3 $R^2$ = .52; Model 4 $R^2$ = .51).

Results from the models using item parcels were again similar (see Figures 2 & 3). All item parcels showed acceptable loadings on their respective factors. The Gv model and the individual Gv abilities models demonstrated acceptable fit (Table 4). Unlike the individual item models, the functionally similar broad Gv model did not indicate that broad Gv was more important in predicting latent APM than reasoning ($\gamma$ = .44 and .38).

**SPM Sample.** Several items, including all Set A of the SPM were not included in analysis due to the low variance of these items. All other items demonstrated significant and substantial loadings on their respective factors.

Five models were tested in this sample; one in which a latent broad Gv factor predicted the residual of the latent SPM variable; one in which each Gv ability independently but simultaneously predicted the residual of the SPM latent variable; and three in which each individual Gv ability independently predicted the residual of the SPM latent variable. The first model was a test of the extent to which broad Gv influences SPM, while the remainder allowed consideration of which of the three narrow abilities (visualization, rotation, flexibility of closure) might demonstrate the strongest prediction of SPM.

The measurement model for Model 1 consisted of latent variables for SPM, inductive reasoning (defined by CAB-I items), flexibility of closure (CAB-Cf items), rotation (MRT items) and visualization (Paper Folding items). The flexibility of closure, mental rotation and visualization factors then defined a broad Gv factor. Model 2 consisted of latent SPM and all latent Gv abilities, without the broad Gv factor. Models 3 – 5 consisted of latent SPM and flexibility of closure (Model 3), rotation (Model 4) or visualization (Model 5). Table 5 contains the fit indices for these models. CFI values were lower than the recommended ≥ .96 and the SRMR values were higher than the recommended ≤ .07, but, given the acceptable chi-square and RMSEA values, and CFI values ≥ .90, examination of structural parameters proceeded.

Table 5

*Fit Statistics for Measurement Models: SPM Sample*

| Model | $\chi^2$ *(df)* | Normed $\chi^2$ | RMSEA | CFI | SRMR |
|---|---|---|---|---|---|
| Individual Items | | | | | |
| Model 1: Broad Gv | 3132 *(2478)* | 1.26 | .03 | .93 | .12 |
| Model 2: All narrow Gv abilities | 3131 *(2474)* | 1.27 | .03 | .93 | .12 |
| Model 3: Flexibility of Closure | 1762 *(1124)* | 1.57 | .04 | .92 | .13 |
| Model 4: Mental Rotation | 1724 *(1077)* | 1.60 | .05 | .90 | .13 |
| Model 5: Visualization | 1998 *(1322)* | 1.51 | .04 | .90 | .13 |
| Item Parcels | | | | | |
| Model 1: Broad Gv | 79.38 *(84)* | 0.95 | < .01 | 1.00 | .03 |
| Model 2: All narrow Gv abilities | 74.34 *(80)* | 0.93 | < .01 | 1.00 | .03 |
| Model 3: Flexibility of Closure | 28.66 *(24)* | 1.19 | .03 | .99 | .02 |
| Model 4: Mental Rotation | 18.20 *(24)* | 0.76 | < .01 | 1.00 | .02 |
| Model 5: Visualization | 14.86 *(24)* | 0.62 | < .01 | 1.00 | .02 |

Examination of Model 1 showed that the latent Gv factor significantly predicted residual SPM performance ($R^2$ = .56, $p$ < .001; Figure 4). In Model 2 (Figure 5) flexibility of closure was the only Gv ability that significantly predicted residual SPM but, together, the three predicted approximately a third of variance in residual latent SPM ($R^2$ = .31, $p$ < .001). Inspection of Models 3-5 showed that flexibility of closure best explained residual latent SPM ($R^2$ = .22, $p$ < .001), followed by mental rotation ($R^2$ = .15, $p$ = .028) and visualization ($R^2$ = .07, $p$ = .106). Again, none of these abilities was able to predict residual latent SPM as well as the latent broad Gv factor.

In Models 1 and 2, broad Gv and the three individual Gv abilities predicted well the variance in SPM performance that was not accounted for by inductive reasoning, although broad Gv accounted for this variance better than the combined individual Gv abilities. A functionally similar model where latent SPM was regressed on both inductive reasoning and latent broad Gv simultaneously indicated that broad Gv may be more important in predicting latent SPM than reasoning (γ = .58 and .33, respectively); however the other four models showed the opposite result: Model 2 (γ = .24 [Flexibility of Closure] and γ = .52 [Inductive Reasoning]); Model 3 (γ = .31 [Flexibility of Closure] and γ = .59 [Inductive Reasoning]); Model 4 (γ = .25 [Mental Rotation] and γ = .63 [Inductive Reasoning]); and Model 5 (γ = .17 [Visualization] and γ = .69 [Inductive Reasoning]). The models including broad Gv and all narrow Gv abilities explained more variance in latent SPM ($R^2$ = .74 and .68, respectively) than did the model including inductive reasoning only ($R^2$ = .60), while the difference was smaller in those models containing only a single Gv ability (Model 3 $R^2$ = .66; Model 4 $R^2$ = .64; Model 5 $R^2$ = .63).

Results from the models using item parcels were again similar (Figures 4 & 5). Both the Gv model and the individual Gv abilities model demonstrated acceptable fit

(Table 5). As with the individual item models, the functionally similar broad Gv model indicated that broad Gv may be more important in predicting latent APM than reasoning, however the remaining models showed the opposite result.

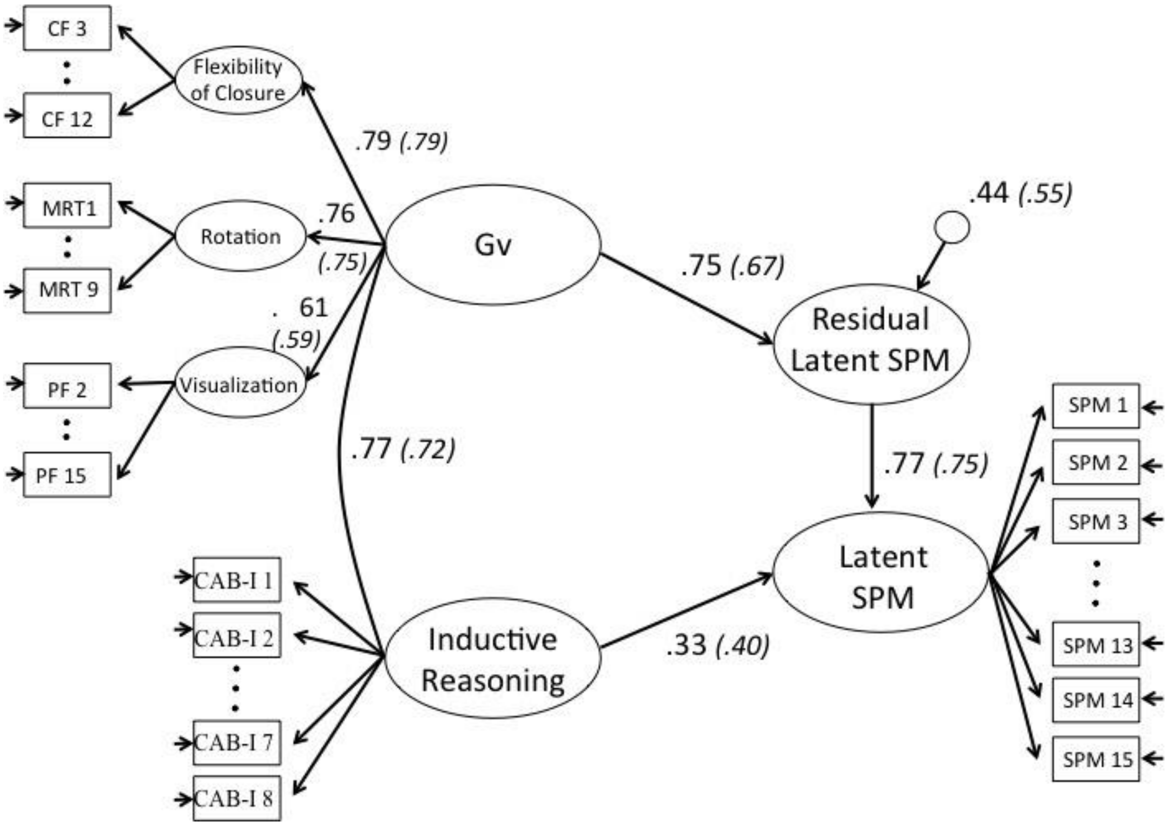


*Figure 4.* SPM Model 1: Structural equation model showing standardized coefficients for latent Gv and inductive reasoning to latent SPM and residual latent SPM. Values in parentheses indicate parameters when item parceling was used.

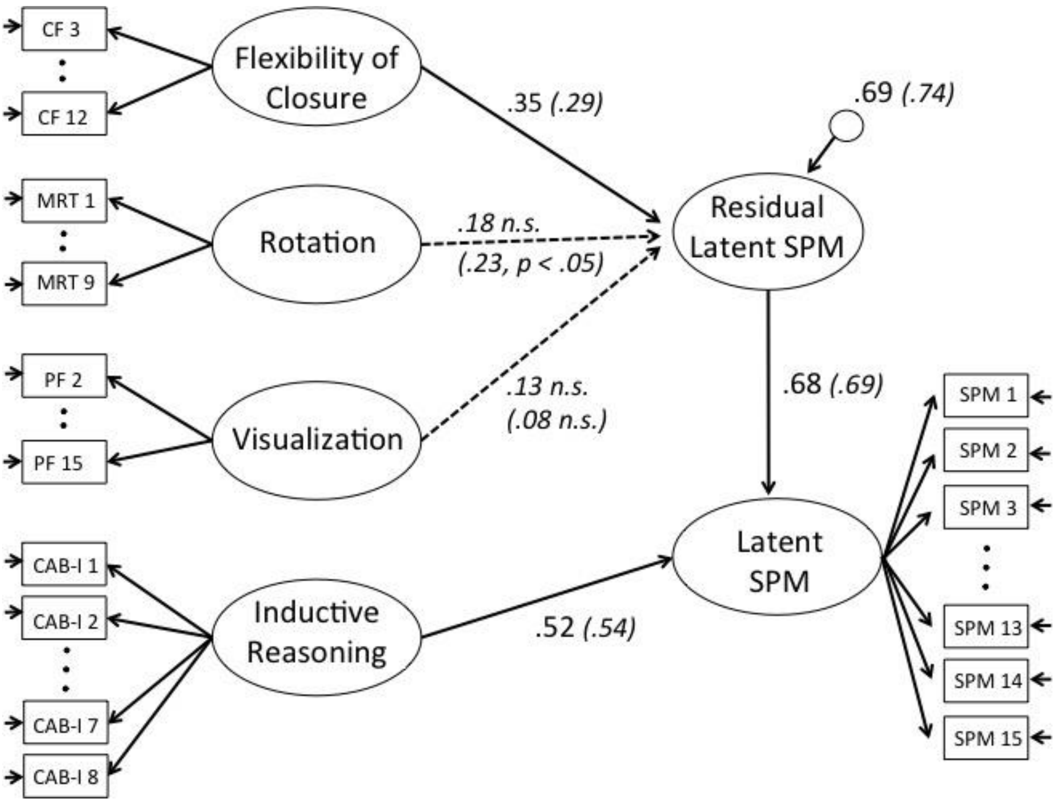


*Figure 5.* SPM Model 2: Structural equation model showing standardized coefficients for ability measures to latent SPM and residual latent SPM. Correlations between factors were modeled but not depicted here for diagram clarity. Values in parentheses indicate parameters when item parceling was used.

## Sex Differences

To explore differences in the relationship between Gv, Gf and APM across males and females, the model described in section 3.3.1 (APM Sample 1) was tested for measurement invariance across sex. This testing was not performed for APM Sample 2 due to the small number of males in that sample, and estimation problems encountered when testing was attempted.

**Multiple groups CFA.** After exclusion of six items due to low variance in the male group, a model with the same pattern of factor loadings showed an acceptable fit in the female ($\chi^2_{(626)}$ = 655, *p* = .202, [normed chi-square = 1.05], RMSEA = .01, CFI = .98, SRMR = .09) and male ($\chi^2_{(626)}$ = 667, *p* = .124, [normed chi-square = 1.07], RMSEA = .02, CFI = .97, SRMR = .13) groups, although the SRMR value for the male group was high.

A sex-invariant measurement model was achieved by freeing the means of rotation and reasoning, as well as APM in males (Table 7). Once this model was established, the regression parameters were added. Again, a significant difference in model fit when these parameters are constrained equal across groups indicates a violation of structural invariance.

Constraining these parameters equal across groups did not result in a significantly worse fit (Table 6). The p-value indicated the difference was not significant and there was no decrease in CFI or increase in RMSEA to indicate that constraining the structural parameters equal resulted in a substantive decrease in model fit. Therefore the relationship was deemed invariant across sex, indicating that sex did not moderate the relationship between Gv, Gf and APM.

Table 6

*Measurement invariance statistics*

| | $\chi^2$ | df | RMSEA | CFI | Δ df | Δ p |
|---|---|---|---|---|---|---|
| Baseline | 1324 | 1252 | .02 | .98 | | |
| Model 2 | 1416 | 1286 | .02 | .96 | 34 | < .001 |
| Mosel 2.2[1] | 1395 | 1285 | .02 | .96 | 33 | < .001 |
| Model 2.3[2] | 1374 | 1284 | .02 | .97 | 32 | .005 |
| Model 2.4[3] | 1360 | 1283 | .02 | .97 | 31 | .135 |
| Model 3[4] | 1360 | 1283 | .02 | .97 | | |
| Model 4 | 1362 | 1285 | .02 | .97 | 2 | .268 |

[1] mean of rotations freed; [2] mean of reasoning freed; [3] mean of APM freed; [4] structural parameters added

**Multiple indicator, multiple cause (MIMIC) model**. MIMIC modeling was used to investigate whether the sex difference in Gv could account for the sex difference in latent APM. The model in which only APM was regressed on sex demonstrated a significant effect of sex (γ = .14, *p* = .025). When the sex difference in Gv was modeled (γ = .24, *p* < .001), sex was no longer a significant predictor of latent APM (γ = .004, *p* = .939), indicating that the sex difference in Gv may be responsible for the sex difference in latent APM.

## Discussion

The aim of this paper was to investigate the role of Gv in RPM performance, and sex differences in the relationship between Gv, Gf and RPM. Results indicated the RPM do engage Gv processes. Gv accounted well for the variance in latent RPM not accounted for by Gf, explaining over half of this residual variance in APM Sample 1 and the SPM Sample. In APM Sample 2 Gv explained slightly less variance in residual latent APM; however, the explanatory power of Gv was still substantial, accounting for up to one-third of the residual variance. In all samples the unique contribution of the broad Gv factor (or rotation in APM Sample 1) to latent RPM was greater than that of Gf. Furthermore, results indicated that the male advantage on Gv could account for the sex difference in latent APM, presenting problems for the interpretation of sex differences in the RPM as reflective of differences in general intelligence.

The results reported here differ from Schweizer et al. (2007), who found that Gv abilities did not improve prediction of latent APM after accounting for reasoning. One major difference between the analyses was the use of item parcels by Schweizer et al. However, a comparison of the results of analyses conducted with and without item parcels in the present paper indicated that this may not have influenced the conclusions as suspected. Comparison of the regression parameters in the item-parceled and non-item-parceled analyses showed consistent results.

The sample composition used in the present study as compared to Schweizer et al. (2007) is unlikely to have influenced results; the samples were very similar, and results indicating invariance of structural parameters across sex suggest differences in the sex composition of the sample should not influence results. Therefore, differences in measures used and the large correlations between Gv and Gf measures in the present study are the most likely sources of the different results.

With regard to the APM, Schweizer et al. (2007) used all items of Set II, while the present study used short forms containing items from Set II. There is evidence for a position effect in the APM (Schweizer, Schreiner, & Gold, 2009), which may be influenced by learning (Ren, Wang, Altmeyer, & Schweizer, 2014). The opportunity for learning within the short forms is necessarily decreased and could have altered the construct measured; a greater involvement of learning may decrease the variance explained by Gv. Furthermore, research suggests while both executive and perceptual attention are related to the position-specific component of the APM, only perceptual attention is related to the ability-specific component (Ren, Goldhammer, Moosbrugger, & Schweizer, 2012). By eliminating some of the position effect in using short forms, this may have increased the role of perceptual attention, which, although not the same as Gv, could arguably contribute to a stronger relationship between Gv and APM in the present case.

There were also differences in the specific Gf ability considered by Schweizer et al. (2007) as compared to APM Sample 1, but not APM Sample 2; a measure of inductive reasoning was used in Schweizer et al. and APM Sample 2, while a combined Gf- Gc measure was used in APM Sample 1. It could be argued that there would be a stronger relationship between the APM and Gf in Schweizer et al.'s study, and in APM Sample 2, than in APM Sample 1, possibly causing Gv to appear more important than is actually the case in APM Sample 1. However, in APM Sample 2 Gv still accounted for a substantial proportion of residual latent APM. In any case, whether or not the Gf measure is specifically an inductive reasoning measure, it should still have a stronger relation to APM than Gv if Gv is not substantially involved in RPM performance.

In the present study the functionally equivalent models demonstrated that Gv may be the main source of performance variance in the RPM, with the regression weights of Gv exceeding those of Gf. This was unexpected and highlights one issue with regard to the analysis and interpretation. As can be seen in the descriptive statistics, variance in Gv measures was larger than in Gf measures and correlations between Gv and Gf measures were larger than expected (and larger than those reported by Schweizer et al. [2007]), indicating substantial overlap of the predictors. Furthermore, in APM Sample 2 and the SPM sample, multiple measures of Gv but only one alternate measure of Gf were used. The greater variance in Gv measures, overlap of predictors, and use of more Gv than Gf measures may have contributed to the dominance of Gv in the present case.

One reason for the higher correlations between measures in the present study could be found in the manner of test administration. In Sample 1, participants completed the measures online, and untimed. This may inflate the correlations between measures due to shared method variance. However, Sample 2 completed the measures in the traditional timed paper-and-pencil form, and the correlations between measures were also larger than those reported by Schweizer et al (2007).

The SPM results, on the other hand, may not be directly comparable to Schweizer et al.’s (2007) study: the SPM may be more likely to be multidimensional than the APM (Colom & Abad, 2007), and therefore also more likely to involve Gv. It is possible that as the items become more challenging, a key distinction between the SPM and APM, a greater analytic ability is required. This idea is apparent in DeShon et al.’s (1995) solution rules for the APM, where visual rules tend to occur at the beginning of the test and analytic rules towards the end. Furthermore, the argument for the decrease in position effect due to the use of a short form or shared method variance due to computerized administration as an explanation for the stronger correlations does not apply to the SPM sample. It should be noted that compared with APM Sample 2, in which identical measures were used with the exception of flexibility of closure, the SPM did show a slightly greater contribution of Gv, suggesting there may be a stronger contribution of Gv in this measure.

It is possible that this difference between APM Sample 2 and the SPM Sample occurred due to the omission of flexibility of closure in APM Sample 2; SPM results pointed to flexibility of closure as the best narrow Gv predictor of RPM performance, although the broad Gv factor did account for more variance in residual SPM than did any of the Gv abilities alone or simultaneously. Research indicating visual salience is an important element in the difficulty of the RPM (Meo et al., 2007; Roberts et al. 2000) supports this. It would certainly be worthwhile to examine the contribution of this narrow Gv ability to latent APM.

With regard to the measures of Gv used, rotation was the best candidate for a Gv predictor in APM Sample 1, while in the other samples this narrow ability explained less than half of the variance in residual latent RPM when considered alone. This could reflect a difference in the measures of rotation or inductive reasoning used. It is possible that the PSVT:R, the measure of rotation in APM Sample 1, involved some reasoning ability given its analogical format, which is different to the MRT. This, combined with the use of the Esoteric Analogies test in this sample, which does not measure inductive reasoning specifically, may have contributed to this discrepancy.

Results of the structural invariance analysis indicated that that sex did not moderate the relationships among Gv, Gf and APM, despite some differences in raw score correlations. This could explain why a male advantage is sometimes reported in APM scores; if females tend to score lower on measures of Gv, and the influence of Gv on the APM is the same across the sexes, then this would result in a lower female score. MIMIC modeling showed that the sex difference in Gv could account for the sex difference in latent APM. This is noteworthy, because sex differences on the APM have been interpreted as differences in *g* (Lynn & Irwing, 2004). The present results indicate that such an interpretation may not be valid, although this finding should be confirmed for the full form APM.

It should be noted, however, that in both APM samples males performed better on all measures. This is not an uncommon finding for the APM, particularly in university samples (Irwing & Lynn, 2005; Lynn & Irwing, 2004), nor for the Gv measures. However, this difference was also found on the alternate Gf measures, which may indicate that the males in these samples were of higher overall cognitive ability than the females. Additionally, it should be noted that the APM samples were samples of highly selected individuals. However, the APM versions used in these samples were specifically designed for use with such samples.

Given the conflicting findings regarding the APM results in the present study and in Schweizer et al.'s (2007) study, future research is needed to confirm the relationship between Gv, Gf and the APM. The fact that the relationships between the constructs were consistent across data sets included in this study does support the findings reported here. However, future research should address any differences between these relationships according to the test form used, as well as consider the narrow Gv ability of flexibility of closure. Furthermore, although a male advantage in Gv was found to account for the sex difference in latent APM, this could not be tested for the SPM with the current sample, and future research should investigate this, given the reported sex difference in the SPM (Lynn et al., 2004; Lynn & Irwing, 2004).

## Notes on Paper 2 Measurement Invariance Testing

The test for measurement invariance when using the WLSMV estimator requires the comparison of two models: a less restrictive and more restrictive model. The less restrictive model is the baseline model, in which means in both groups are fixed at zero but factor loadings and thresholds are freely estimated. The more restrictive model constrains factor loadings and thresholds invariant across groups, while fixing the factor means to zero in the first group and freely estimating these means in the second group.

It should therefore be noted that the most relevant test in Table 6 is between the baseline model and model 2.4, which indicates that the difference in fit between these two models (the less and more restrictive) is not significantly different, confirming measurement invariance. These analyses were conducted earlier in candidature, before significant experience with measurement invariance testing was obtained, and therefore the tests of model 2.2 and 2.3 compared to the baseline model contain unnecessary information. There were significant mean differences found for all factors once invariance of loadings and thresholds was confirmed in any case.

Furthermore, the reader may note that the same sample was used for measurement invariance testing in Papers 1 and 2, and that in Paper 1, this sample did not show invariance across sex not only on the mean of APM, but also item 14 of the APM. This was not found in Paper 2 because the model tested for invariance was more complex and consequently the decrease in model fit when constraining this item equal was not sufficient to demonstrate the lack of invariance in this case.

# Chapter 5: Paper 3

This paper aimed to investigate the evidence for sex differences in manifest scores on different measures of inductive reasoning. The rationale for this paper came from literature providing evidence that the sex difference on measures of Gf may vary depending on the specific measure used (Colom & Garcia-Lopez, 2002), and also that sex differences on the RPM are caused by sex differences on Gv (Colom et al., 2004). Two meta-analyses on sex differences on the RPM tests have reported a male advantage (Irwing & Lynn, 2005; Lynn & Irwing, 2004b). However, various criticisms have been proposed regarding certain analytical choices (Blinkhorn, 2005, 2006), and these meta-analyses are now over a decade old. Thus we felt it necessary to first re-address the evidence for a male advantage on the RPM tests. Additionally, Lynn and Irwing (2004b) and Irwing and Lynn (2005) interpret the male advantage on the RPM as evidence for a male advantage on *g*. We therefore thought it pertinent to investigate whether inductive reasoning tests in general, not solely the RPM, demonstrate a male advantage. If that were not the case, that would indicate that it was perhaps something unique about the RPM tests that was causing the male advantage, and thus interpreting a male advantage on the RPM tests as a male advantage in *g*, or even Gf, would be problematic. This study also allowed us to assess how different content facets, as per the BIS model, might influence sex differences in measures of reasoning.

Following is a full description of the search strategies used, the inclusion criteria employed, and an explanation and justification for these choices.

# Methodology

## Selection of Measures to Include

The aim of this meta-analysis was to investigate sex differences in tests of inductive reasoning and whether the existence or magnitude of sex differences may depend on the type of test used. The method section of Paper 3 details the identification process for measures of inductive reasoning.

Forty-one tests were identified for inclusion (see Paper 3, Table 1). An additional 25 tests were identified as measures of inductive reasoning, but excluded because the recommended age range for testing did not extend beyond 18 years. Verbal analogies and classification tests were included, because although they can be viewed as combined measure of Gf/Gc, at their core, the “analogies” and “classification” format requires inductive reasoning. It is acknowledged that analogies presented in the form of worded problems will introduce a degree of Gc; however, it was decided that it would be useful to also investigate how this stimulus type compares to others. Similarly, number series measures were included. Although these tests are arguably more representative of quantitative reasoning than inductive reasoning, they have shown significant loadings on an inductive reasoning factor (Carroll, 1993). It is also acknowledged that there very well may be more tests of inductive reasoning not included in the list of identified tests; nonetheless, it is argued that this list contains a sufficiently broad sample of different types of the most commonly used inductive reasoning tests to provide an adequate test of the questions investigated here.

**Literature Search**

Due to the large amount of cognitive ability literature available, extensive investigation into the most efficient search terms was conducted before deciding on the final search strategy to be used. Although a comprehensive search of all existing published work relevant to the research question would be desirable, this simply was not feasible within the time constraints available for completion of the PhD degree under the current Australian system. Initial searching of potentially relevant papers from all years returned over 50,000 results in PsycINFO alone (reduced to approximately 37,000 when the WAIS was excluded). Therefore, two options were considered; (i) to narrow the range of years included, or (ii) to take a random sample of all years. No strong theoretical rationale was defensible for narrowing the year range and, in any case, it was apparent that the vast majority of published literature has become available during the two decades. Thus, the decision was made to include a random selection of years.

Using a random selection of years should not introduce any bias into the results because missing studies should not differ systematically from included studies. However, in order to check for bias two separate randomised searches were performed, each including 14 years. A two-step process was followed to select years for inclusion: firstly, a random year was chosen from each decade over the last century (starting at the range 2006-2015 and ending at 1916-1925). Following this, random years from the whole range (1916-2015) were selected until the number of years selected reached 14 (approximately 3,000 search results each; see Table 5.1).

Table 5.1

*Years searched*

| Sample 1 | Sample 2 |
|---|---|
| 2010 | 2009 |
| 1996 | 2005 |
| 1991 | 1995 |
| 1981 | 1977 |
| 1970 | 1975 |
| 1961 | 1965 |
| 1947 | 1954 |
| 1946 | 1939 |
| 1934 | 1932 |
| 1923 | 1918 |
| 2008 | 1934 |
| 1973 | 1990 |
| 1931 | 1971 |
| 1942 | 1980 |

A comprehensive search of the PsycInfo, Scopus and ScienceDirect databases was undertaken for the selected years. It is acknowledged that ScienceDirect is a database for Elsevier journals only; however, it has the power to search full-text, which was particularly advantageous when searching for published articles that had utilized a specific test, which is sometimes not listed anywhere aside from in the full-text of the article. The ERIC database was initially also included but it produced very few relevant results, mainly due to the age restrictions of the current study, and searching of ERIC was subsequently discontinued.

## Search Terms

Given the abundant cognitive ability literature, inclusion of both terms “sex differences” and “gender differences” in the searches was initially considered as a strategy for targeting more relevant results as well as reducing the total number of search results. However, subsequent consideration of preliminary results suggested that there was a real potential to bias results by including these search terms. The intention was to find as many papers as possible that reported the existence or absence of a sex difference on the relevant measures without this necessarily being the purpose of the paper. This is because studies with significant results or larger effect sizes are more likely to be published; however, including research where it was not the explicit intention of the study to examine sex differences in the intelligence measures should reduce any publication bias apparent in the results. However, it is acknowledged that this may not completely remove bias, because the existence of sex differences may be more likely to be reported than the absence of sex differences, even if that was not the goal of that particular article.

The inclusion of general search terms such as “cognitive ability”, “fluid ability” etc., as subject headings and in titles and abstracts was considered. For a search of the randomly chosen years 1983 and 2011 in PsycINFO, using only test names detected 22 potentially relevant papers; and including the general search terms returned only one additional article that met inclusion criteria beyond those results obtained by using the test names only. Moreover, inclusion of general search terms substantially increased the total number of search results (from 30,042 to 50,208 across all years). Therefore the final searches included only the relevant test names (in all fields; see Paper 3: Table 1). Similarly, with Scopus, it was found that general search terms of title, abstract and keywords identified zero papers that met inclusion criteria out of the first 200 identified for the year 2011, while the test names only search produced four out of the first 100. For the ScienceDirect search, the full-text of the article is searched for the search terms, and thus it was not necessary to include any more general terms other than the test names.

Finally, the inclusion of the term “Wechsler Adult Intelligence Scale” was found to be problematic. This term was responsible for nearly 20,000 of the 30,042 results obtained in the PsycINFO search and it was noted that the vast majority of papers identified using this search term were concerned with clinical or otherwise “non-normal” populations. A search for the WAIS in the year 2010 identified only one paper out of 1947 meeting criteria and therefore the WAIS subtests were not included in the search terms. They have been, however, included in the meta-analysis where the relevant information was retrievable from papers identified through other search terms.

### Sample Three

An additional search was performed to gather data for a third sample. This sample is comprised of data extracted from papers authored by researchers identified as prolific publishers in the area of intelligence testing, or researchers identified as having published cognitive ability data obtained from large population-representative samples. A search for all published papers listing these individuals as authors was conducted using PsycINFO, Scopus and ScienceDirect. These authors were then contacted to request clarification on any data not reported in papers identified as using the desired measures in a sample that met the inclusion criteria. This was accomplished with varying levels of success. It should be noted that there may be some bias in the results obtained from this sample, because many of the papers included sought specifically to investigate sex differences.

Several papers identified in Sample 3 were also identified in either Sample 1 or Sample 2. In these cases, they were included in Sample 1 or 2 and not included in Sample 3, with the exception of data requests from authors. Where papers were identified in Sample 1 or 2, but information was obtained from the author, they were included in Sample 3.

## Grey Literature

Grey literature was not included in this meta-analysis, with the one exception of standardisation data. Grey literature such as book chapters, conference papers and theses were not included, partially due to the necessity to keep the number of results to a manageable level, but also as a form of quality control. Standardisation data were included, however, because standardisation samples often consist of the most population-representative data available for cognitive ability tests and therefore conclusions drawn from such data would be more compelling. Inclusion of standardisation data should also reduce the problem of publication bias in the identified published literature.

To begin, technical manuals were consulted to determine whether they provided test or subtest totals by sex, for the appropriate age range. Where they did not, test publishers were contacted to request the data. These attempts were largely unsuccessful, with the exception of the General Reasoning Test 2, for which a large amount of Australian data was provided; and these have been included. Some further results from standardisation data were obtained through the regular search of published papers, and have been included. In some cases where standardisation data could not be obtained for the relevant age range only, these data were included anyway due to the high quality information provided (e.g. data for the K-BIT Matrices concerning the age range 18-90 has been included) Similarly, where published studies reported data concerning standardisation samples which violated the age range, these samples were included if the majority of the sample was within the age range.

**Reference Lists of Identified Papers**

The usual meta-analytic strategy of identifying further papers from the reference list of already identified papers was not used in this case. This was because it is not immediately obvious which references may report data relating to sex differences in tests of inductive reasoning, and it was not feasible to inspect every article in the reference list.

Further to this, there have been a handful of previous meta-analyses concerning some of the tests included in this study; particularly the RPM. Data from these meta-analyses were not included in this study so as to preserve the strategy of random sampling of data. Likewise, data contained in review papers was not included.

**Data Collection**

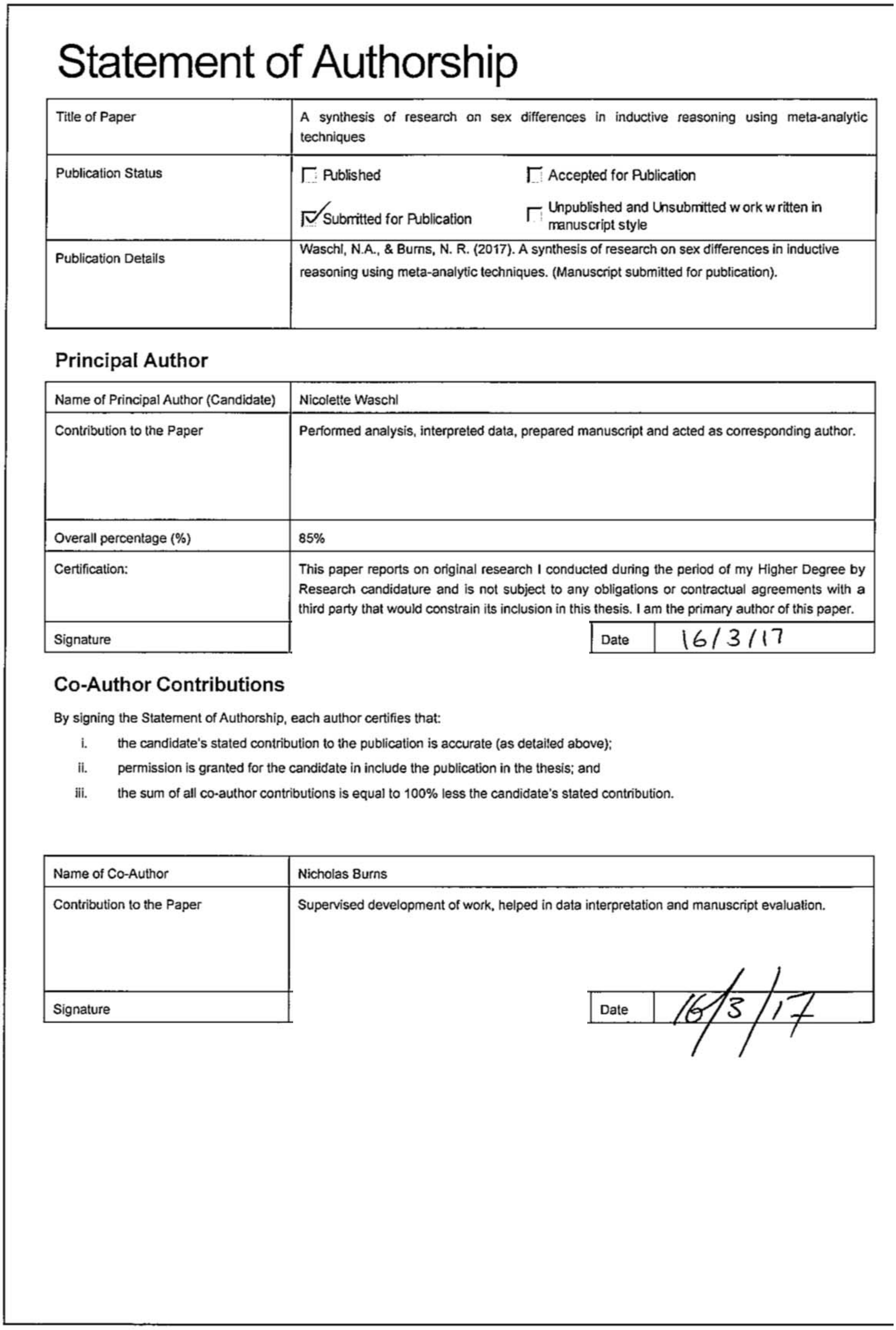
Literature searches were conducted between the period of October 2015 and May 2016 (Sample 1: October 2015-December 2015; Sample 2: January 2016-March 2016; Sample 3: March 2016-May 2016). The literature search initially identified 8,437 unique articles. PRISMA diagrams (Appendix C) provide information regarding search numbers. A description of the inclusion criteria for this study can be found in the method section of Paper 3.

# Paper 3

# A synthesis of research on sex differences in inductive reasoning using meta-analytic techniques

Waschl, N.A., & Burns, N. R. (2017). A synthesis of research on sex differences in inductive reasoning using meta-analytic techniques. (Manuscript submitted for publication).

*Note.* Supplementary materials can be found in Appendix C



# Statement of Authorship

| | |
|---|---|
| Title of Paper | A synthesis of research on sex differences in inductive reasoning using meta-analytic techniques |
| Publication Status | ☐ Published ☐ Accepted for Publication<br>☑ Submitted for Publication ☐ Unpublished and Unsubmitted w ork w ritten in manuscript style |
| Publication Details | Waschl, N.A., & Burns, N. R. (2017). A synthesis of research on sex differences in inductive reasoning using meta-analytic techniques. (Manuscript submitted for publication). |

## Principal Author

| | | | |
|---|---|---|---|
| Name of Principal Author (Candidate) | Nicolette Waschl | | |
| Contribution to the Paper | Performed analysis, interpreted data, prepared manuscript and acted as corresponding author. | | |
| Overall percentage (%) | 85% | | |
| Certification: | This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper. | | |
| Signature | | Date | 16/3/17 |

## Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

| | | | |
|---|---|---|---|
| Name of Co-Author | Nicholas Burns | | |
| Contribution to the Paper | Supervised development of work, helped in data interpretation and manuscript evaluation. | | |
| Signature | | Date | 16/3/17 |

## Abstract

Meta-analyses concerning sex differences in measures of fluid ability have been limited to discussions of sex differences in the Raven's Progressive Matrices. This study synthesized data concerning sex differences on several different types of inductive reasoning tests in order to assess the evidence for an overall sex difference in manifest scores, as well as whether the magnitude of the sex difference may vary depending on certain test characteristics such as stimuli and item type. Meta-analytic techniques are used to summarize data concerning sex differences in inductive reasoning from 94 studies. The summary effect sizes indicate that there is significant variation in the magnitude, and in some cases direction, of the effect size in measures of inductive reasoning. Further investigation is needed in order to determine the cause of this variation; however, implications for the variation in the effect size of the sex difference across tests is discussed in light of theoretical interpretations and practical uses.

*Keywords*: sex differences; fluid ability; inductive reasoning; meta-analysis

The study of sex differences in intelligence has received a lot of attention over the years, with current consensus suggesting that there is no sex difference in general intelligence (Calvin, Fernandes, Smith, Visscher & Deary, 2010; Camarata & Woodcock, 2006; Keith, Reynolds, Roberts, Winter & Austin, 2011; Mackintosh, 1996), but that there may be sex differences in specific abilities, such as a male advantage on certain forms of spatial ability (Keith et al., 2011; Keith, Reynolds, Patel & Ridley, 2008; Linn & Petersen, 1985; Reynolds, Keith, Ridley & Patel, 2008; Voyer, Voyer & Bryden, 1995) and a female advantage in processing speed (Camarata & Woodcock, 2006; Keith et al., 2008; Keith et al., 2011). Other meta-analyses indicate little sex difference in verbal ability (Hyde & Linn, 1988), but a weak to moderate female advantage for school marks in language subjects (Voyer & Voyer, 2014). Recent research also indicates little sex difference in mathematic ability and mathematic school marks (Lindberg, Hyde, Petersen & Linn, 2010; Voyer & Voyer, 2014). One question that has not yet been satisfactorily answered concerns sex differences in manifest scores on measures of fluid ability (Gf), and specifically inductive reasoning, the ability to discover the underlying rules governing a given relationship or phenomenon, and apply those rules to obtain a solution (Carroll, 1993). Due to the fact that measures of inductive reasoning, for example the Raven’s Progressive Matrices (RPM) and other figural matrices tasks, are often used in practice as a sole measure to represent general intelligence, an understanding of sex differences in scores on such measures is vital.

## Theoretical Framework

Several models of the structure of cognitive abilities exist; one with arguably the most empirical and theoretical support (Ortiz, 2015) is the Cattell-Horn-Carroll (CHC) model, based on Cattell and Horn’s Gf/Gc theory (Horn, 1965; 1994) and Carroll’s (1993) three-stratum theory. This model is hierarchical, consisting of narrow abilities, broad abilities and a general intelligence factor. Under the CHC model, fluid ability represents a broad factor, subsuming the narrow abilities of inductive, general sequential (deductive) and quantitative reasoning. The most representative of Gf abilities is inductive reasoning (Carroll, 1993), also arguably the most commonly used construct in assessing Gf. Inductive reasoning is typically measured using figural tasks, due to the need to present item stimuli that are either equally unfamiliar or equally familiar to all test-takers. However, other types of inductive reasoning tasks do exist (e.g. letter series, which uses alphabetic stimuli).

The question of sex differences in Gf, and the interpretation of its result, must also confront the critically important theoretical issue of whether or not Gf is equivalent to *g*, or general intelligence. Some argue that these constructs are statistically indistinguishable, and therefore equivalent (Gustafsson, 1984), while others argue for the validity of the Gf-*g* distinction (Carroll, 2003; Gignac, 2007). If we subscribe to the notion that Gf and *g* are equivalent, then this can lead to the interpretation of any sex differences found on measures of Gf as differences in *g*.

An example of this issue can be extended to the Raven’s Progressive Matrices (RPM) set of tests. These tests were developed to measure *g* and are often still used and interpreted as such. Although these tests may have one of the highest *g*-loadings of available measures (Jensen, 1998), under the CHC model of intelligence, these tests can be more appropriately considered a measure of inductive reasoning at the narrow ability level, and Gf at the broad level. Even though current consensus suggests there is no sex difference in general intelligence, sex differences on the RPM tests are sometimes interpreted as sex differences in *g*. We must be careful in making such overarching conclusions, particularly from a single measure and from manifest scores only. Therefore, in the present case we consider Gf at the broad level, and inductive reasoning at the narrow level, to be separate from *g*, and make no conclusions regarding sex differences in general intelligence.

## Previous Meta-Analyses

Putatively, the first meta-analyses of sex differences in fluid ability were carried out by Lynn and Irwing (2004b) and Irwing and Lynn (2005); however, these authors considered the RPM tests only. Both meta-analyses reported a male advantage on the RPM tests and interpreted this as indicating superior general intelligence in males, to the tune of $d$ = .33, or nearly 5 IQ points. This is in contrast to an earlier, qualitative review of sex differences in the RPM tests that concluded that there was no significant difference (Court, 1983).

Blinkhorn (2005, 2006) has criticized Lynn and Irwing (2004b) and Irwing and Lynn's (2005) work. Criticisms include the exclusion of a particularly large study that would have substantially altered results, not weighting results by sample size, and overestimating the size of the sex difference reported in the IQ metric. Lynn and Irwing (2004b) are additionally criticized for presenting very few meta-analytic statistics, and relying only on reporting effect size estimates. Other criticisms regarding sample selectivity and issues such as the file-drawer problem are also presented. As maintained by Irwing and Lynn (2006), in a case such as this the file-drawer problem is minimized through the inclusion of studies not expressly designed to investigate sex differences. The issue of sample selectivity is harder to overcome; many large, representative datasets are often difficult to obtain. In the present case, attempts were made to access such data, but representativeness of the samples was also considered in analysis.

One issue Lynn and Irwing (2004b) raise with regard to the meta-analysis of sex differences in abilities is the "apples and oranges" problem; that to investigate sex differences in abilities we cannot pool different phenomena and expect a consistent effect. It is for this reason that they chose to investigate a single ability (which they refer to as reasoning), and specifically the RPM tests only. However, studying a single test may be too limiting to allow conclusions regarding sex differences in reasoning, let alone in *g*. Although the RPM tests are widely considered an excellent measure of reasoning, using only one test to make conclusions about a certain cognitive ability can be problematic due to test specificity. Recent research has suggested that up to 25% of variance in the RPM tests may be test specific (Gignac, 2015).

One example of this issue with respect to the RPM tests is that item solution may require a degree of spatial ability (e.g. Colom, Escorial & Rebollo, 2004; DeShon, Chan & Weissbein, 1995). Given the robust finding that males show an advantage on tests of spatial ability (Linn & Petersen, 1985; Voyer et al., 1995), interpreting a male advantage on the RPM tests as evidence of a male advantage in inductive reasoning or, indeed, *g*, can be considered problematic. The issue of the relationship between Gf and Gv is an ongoing one in the literature. Many tests of Gf, particularly of inductive reasoning, utilize figural stimuli. Indeed, Colom et al. (2004) report findings indicating that spatial ability may be responsible for the male advantage identified on the RPM tests and it is possible that the stimuli used could be responsible for this.

## Moderators

Given suggestions that the figural stimuli used in many inductive reasoning tasks may lead to the involvement of spatial ability in these tasks and therefore a male advantage, one proposed moderator of sex differences in inductive reasoning at the test level is the item stimuli. Research under the framework of the Berlin model of Intelligence Structure (Jäger, 1982; Süβ & Beauducel, 2005) provides evidence that the "content" factor (e.g. figural, verbal, numerical) plays a role in the ability measured; although the operation facet (the cognitive processes involved, e.g. reasoning, memory; likened to the second-order CHC factors such as Gf, Gs [processing speed]) has been found to account for more variance in measures of intelligence than the content facet, models accounting for both facets have demonstrated superior fit to the data (Bucik & Neubauer, 1996). In the case of inductive reasoning tasks, it is proposed that the figural content may be a moderator of the male advantage often reported.

The specific type of task used to measure a given ability could also conceivably influence any sex differences identified. Inductive reasoning tasks come in many different types, including: matrices, similarities, series, and classification (Carroll, 1993). Previous meta-analyses of other narrow cognitive abilities have found that the magnitude of the sex difference can depend on the type of task under consideration. For example, with regard to visuospatial abilities, mental rotation tests show the largest male advantage, while measures of visualization tend to show the smallest difference (Linn & Petersen, 1985; Voyer et al., 1995).

Studies that have administered several different measures of inductive reasoning to the same sample have reported different magnitudes and directions of the sex difference depending on the particular measure. For example, Burgaleta et al. (2012) found a slight female advantage on the Advanced RPM (APM; *d* = -0.11), a larger female advantage on the Primary Mental Abilities – Reasoning (PMA-R; *d* = -0.55) and a slight male advantage on the Differential Aptitude Test – Reasoning (DAT-AR; *d* = 0.11)[1]. Similarly, Colom and Garcia-Lopez (2002) found a slight male advantage on the APM (*d* = .28), a slight female

[1] A negative effect size indicates a female advantage and a positive effect size indicates a male advantage. This convention is used throughout this paper.

advantage on the PMA-R (*d* = -0.19) and no substantial difference on the overall score obtained from Cattell’s Culture Fair Intelligence Test (CFIT; *d* = 0.10). The APM and DAT-AR use figural matrices and figural series tasks, respectively, while the CFIT combines matrices, series, classification and similarities, all in figural format, and the PMA-R uses alphabetic series tasks.

**Present Study**

Given criticisms of the previous meta-analyses on sex differences in the RPM tests, and that these analyses were completed over a decade ago, the first aim of the current study was to review and update the evidence for a sex difference in these tests. Second, this research synthesis sought to investigate sex differences within the specific narrow ability of inductive reasoning, but at a broader level than just the RPM tests, and therefore to determine what the evidence is for an overall sex difference in inductive reasoning. Since other research utilizing different tests of inductive reasoning suggests that these other tests may not demonstrate the sex difference sometimes found in the RPM tests (e.g. Colom & Garcia-Lopez, 2002), this study additionally sought to investigate whether the sex differences reported in the RPM tests are consistently found in other tests of inductive reasoning. One hypothesized reason for the sex difference in the RPM tests specifically is the item format, namely abstract figural matrices. Therefore, sex differences were investigated with regard to the test characteristics (stimuli and question type) as a moderator.

# Method

## Selection of Measures to Include

In order to determine which tests of inductive reasoning to include in the current meta-analysis, several sources were consulted, including Carroll's (1993) *Human Cognitive Abilities: A Survey of Factor Analytic Studies*, which provided the definition of inductive reasoning adopted here and descriptions of the types of tasks found to be good measures of this ability. Murphy, Spies and Plake's (2006) *Tests in Print VII* was also consulted, as was McGrew and Flanagan's (1998) *The Intelligence Test Desk Reference*. Additionally, tests available in the University of Adelaide Psychology Test Library were inspected. Verbal analogies and classification tests were also included because, although they are not typically considered as measures of inductive reasoning but rather measures of both Gf and Gc, at their core, the "analogies" and "classification" formats require inductive reasoning.

Forty-one tests were identified for inclusion (see Table 1). These tests were determined to be the most commonly used, age appropriate measures of the desired construct, inductive reasoning. It is acknowledged that there very well may be additional tests of inductive reasoning not included in the list of identified tests; nonetheless, it is argued that this list contains a sufficiently broad sample of different types of the most commonly used inductive reasoning tests to provide an adequate test of the questions investigated here.

Table 1

*Measures of inductive reasoning identified and included in searches*

| Test Name | Test Battery | Description | Type | Stimuli | Papers Identified |
|---|---|---|---|---|---|
| Inductive Reasoning | CAB | Identify the pattern in a sequence of letter sets, and determine which set does not belong. | Classification | Letters | Yes |
| Classification | CFIT | Select which figure is different from the other four (Scale 2) or which two are different from the other three (Scale 3). | Classification | Figural | Yes |
| Conditions | CFIT | Select the figure which duplicates the conditions given in the example. | Classification | Figural | Yes |
| Matrices | CFIT | Correctly complete the figural design. | Matrices | Figural | Yes |
| Series | CFIT | Select the answer that best continues the series of figures. | Series | Figural | Yes |
| Abstract Reasoning | DAT | Determine which alternative comes next in a series of abstract figures. | Series | Figural | Yes |
| Verbal reasoning | DAT | Complete a sentence stated in the form of an analogy. The first and last words from the sentence are missing, and a pair of words must be selected to complete the sentence. | Analogies | Verbal | Yes |

| Test Name | Test Battery | Description | Type | Stimuli | Papers Identified |
|---|---|---|---|---|---|
| Figure Classification Test | ETS | Identify which group (out of two or three groups) a figure belongs to by identifying the rule governing group membership. | Classification | Figural | No |
| Letter Sets | ETS | Identify the pattern in sequences of letter sets, and determine which sequence does not belong. | Classification | Letters | Yes |
| Locations test | ETS | Identify the rule governing the location of an “x” in a given group of rows of dashes, and apply the rule to determine where the “x” belongs in the next row. | Rule Discovery | Visual (Basic) | No |
| Abstract Reasoning | GRT2 | Figural analogies, similarities and series. | Various | Figural | Yes |
| Numeric Reasoning | GRT2 | Numeric analogies, series and word problems. | Various | Numeric | Yes |
| Verbal Reasoning | GRT2 | Verbal analogies, similarities, classification, word meanings. | Various | Words | Yes |
| Matrices | IST | Figures are arranged according to a particular rule. The task is to choose which figure from the answer options conforms to this rule. | Classification | Figural | No |
| Number Sequences | IST | A series of numbers formed according to a specific rule are presented. The task is to find the next number in the series. | Series | Numeric | No |

| Test Name | Test Battery | Description | Type | Stimuli | Papers Identified |
|---|---|---|---|---|---|
| Verbal analogies | IST | Identify the relationship between two words and apply the rule governing the relationship by choosing a word that shows a similar relationship to another given word. | Analogies | Verbal | No |
| Verbal similarities | IST | The task is to choose those two words, from a possible six, for which there is a common collective term. | Classification | Verbal | No |
| Matrices | K-BIT | Solve either a 2x2 matrix, a 3x3 matrix, or complete a pattern of dots. | Matrices | Figural | Yes |
| Four-letter words | K-SNAP | The test-taker is presented with clues involving a series of four-letter words and is required to discover "secret" words by studying these clues. | Misc. | Verbal | No |
| Mystery Codes | KAIT | The test-taker is presented with codes associated with a set of pictorial stimuli. The task is then to figure out the code for a novel stimulus. | Misc. | Pictorial | Yes |
| Reasoning | PMA | Discover the rules governing a series of letters and mark the letter that should come next in the series. | Series | Letters | Yes |
| APM | RPM | Identify the missing element of a figural matrix. | Matrices | Figural | Yes |

| Test Name | Test Battery | Description | Type | Stimuli | Papers Identified |
|---|---|---|---|---|---|
| SPM | RPM | Identify the missing element of a figural matrix. | Matrices | Figural | Yes |
| Matrices | S-B IV | Identify the missing element of a figural matrix. | Matrices | Figural | No |
| Word Series | Schaie, 1985 | Parallels PMA reasoning - the participant is shown a series of words (e.g. Jan, March, May) and is asked to identify the next word in the series. | Series | Words | Yes |
| Abstraction | Shipley | Fill in the blanks with the answer that best completes the pattern. | Series | Various | Yes |
| Letter Series | Various | Various tests similar to the PMA-R subtest (including ADEPT letter series and Gf/Gc Quickie Battery letter series). | Series | Letters | Yes |
| Matrix Reasoning | WAIS | Identify the missing element of a figural matrix. | Matrices | Figural | Yes |
| Similarities | WAIS | Describe how two given things are alike. | Classification | Verbal | Yes |
| Concept formation | WJ-III | Identify the rules for concepts when shown illustrations of instances of the concept and non-instances of the concept. | Rule Discovery | Pictorial | Yes |
| Number Matrices | WJ-III | Complete an analogy in the form of a numerical matrix. | Analogies | Numeric | Yes |
| Number Series | WJ-III | Discover and apply the underlying rule of a numerical sequence. | Series | Numeric | Yes |

| Test Name | Test Battery | Description | Type | Stimuli | Papers Identified |
|---|---|---|---|---|---|
| Verbal Analogies | WJ-R | Verbal analogies of the form A:B::C:? | Analogies | Verbal | No |
| Alice Heim 4 | - | Verbal and non-verbal reasoning items in the form of analogies, series, classification etc. | Various | Various | Yes |
| Aros Number Series | - | Identify the mathematical basis of the series and then add the next two numbers in the series. | Series | Numeric | No |
| Horn's Reasoning Scale | - | Identify the number or letter which does not fit the series. | Classification | Numeric / Letters | No |
| Letter Grouping II | - | Decide which group of letters does not belong with the other groups of letters. | Classification | Letters | No |
| Miller Analogies Test | - | Verbal analogies of the form A:B::C:? | Analogies | Verbal | Yes |
| TONI-3 | - | Identify the missing element of a figural matrix. | Matrices | Figural | No |
| USTM Number series | - | Identify the mathematical basis of the series and then add the next number in the series. | Series | Numeric | No |
| Visual Analogies Test | - | Visual analogies of the form A:B::C:? | Analogies | Figural | No |

*Note.* APM = Advanced Progressive Matrices; AH-4 = Alice Heim Test; CAB-I = Comprehensive Ability Battery – Inductive Reasoning; CFIT = Culture Fair Intelligence Test; DAT-AR = Differential Aptitude Test – Abstract Reasoning; DAT-VR: Differential Aptitude Test – Verbal Reasoning; Gf = latent Gf factor; GRT2 = General Reasoning Test (AR = Abstract Reasoning; NR = Numerical Reasoning; VR = Verbal Reasoning); KAIT = Kaufmann Adult Intelligence Test; KBIT = Kaufmann Brief Intelligence Test; MAT = Miller Analogies Test; PMA-R = Primary Mental Abilities – Reasoning; RPM = Raven's Progressive Matrices (Advanced or Standard not specified); SILS = Shipley Institute of Living Scale; SPM = Standard Progressive Matrices; WAIS-MR = Wechsler Adult Intelligence Scale Matrix Reasoning; WAIS-S = Wechsler Adult Intelligence Scale Similarities; WJIII-CF = Woodcock-Johnson III Concept Formation.

## Literature Search

**Sample 1 and Sample 2.** Although meta-analyses typically seek to include all published studies on a certain topic, the topic of sex differences in cognitive abilities faces the problem of an enormous body of literature. Furthermore, as Lynn and Irwing (2004b) and Irwing and Lynn (2005) point out, many studies that report the desired information may not indicate this in their title or abstract. The intention of the current analysis was to not only identify published research with the aim of investigating sex differences in cognitive abilities, but also those papers which report it as an aside. The consequent body of literature to be searched was extremely large. Therefore, a random selection of years from the period of 1915-2015 was searched. Although this is not the typical meta-analytic strategy, the use of random sampling should offset any bias while reducing the search results to a manageable number. To further ensure unbiased results, two searches, each including 14 randomly selected years, were undertaken, with their results to be compared. A systematic search of PsycINFO, Scopus and ScienceDirect was conducted between November 2015 and July 2016. Search terms used were test names (all fields searched; see Table 1)[2]. Searches were limited to English language, peer-reviewed documents only.

[2] WAIS was not included in search terms due to the overwhelming number of results, most of which appeared to concern clinical populations. However, results from this test were included when papers included in the literature search reported relevant results.

**Sample 3.** A common strategy in conducting meta-analyses is to contact researchers who have published extensively in the area; we modified this strategy slightly by conducting a third search of papers published by researchers identified as prolific publishers in the area of intelligence testing, or who had published papers utilizing large, population-representative data sets. In addition, these researchers were contacted to request information when it was not presented in their published work, and to request any unpublished data they may have.

Also included in Sample 3 were data relevant to standardization samples. In the case of ability tests, standardization samples represent among the best, population-representative data available. This information was sought from test manuals, and in cases where it was not reported in test manuals, was requested from publishers or individuals who had published using such data.

## Inclusion Criteria

To be eligible for inclusion in this meta-analysis, a study had to be an original research paper meeting the following criteria: (1) use of a standardized test of inductive reasoning (or verbal analogies or classification), as previously identified in Table 1; (2) participants aged between 18-64.9 years, inclusive; (3) non-clinical population (i.e. no intellectual or physical disabilities, mental or physical illness); (4) no experimental manipulation before completion of the inductive reasoning measure; (5) sample size >99; and (6) report sufficient statistics required to calculate an effect size.

The restriction of age to between the years 18-64.9 was in order to avoid any developmental effects. There is some indication that there is variability in sex differences during development from childhood until approximately the age of 16 (Keith et al., 2008; Lynn, 1999). Therefore, restricting the minimum age to 18 years is a conservative restriction to avoid any developmental effects of childhood. Likewise, the upper age limit of 64.9 years was adopted to avoid any confounding effects of cognitive decline.

In several cases multiple studies were identified which reported results from the same sample on the same measure or measures (sometimes at different time points, sometimes at the same time point). There were small differences in the number of participants included in these studies; therefore data reported from the study are those that used the largest sample. Data from previous meta-analyses and reviews were not included in the present study in order to preserve the random selection of studies. Finally, during the course of the literature search, a paper reporting results for male dentists and female dental assistants was identified. This was excluded based on the fact that it was likely biased towards including males with higher cognitive ability.

## Meta-Analytic Techniques and Procedures

**Effect size calculation.** Effect sizes were calculated using the compute.es (Del Re, 2013) package in R version 3.2.3 (R Core Team, 2015). In the majority of cases, sample sizes, means and standard deviations of male and female groups were used to calculate Hedge's *g* and its associated variance. Where these statistics were not reported, results from t-tests (*t* and sample size), Pearson's *r* (*r* and sample size) or Cohen's *d* (*d* and sample size) were used to calculate Hedge's *g*.

**Analysis.** In order to conduct the meta-analysis, the R package metafor (Viechtbauer, 2010) was used. Package metafor was used to calculate summary effect sizes and associated heterogeneity statistics as well as to conduct moderator analyses. A random-effects model with inverse-variance weighting was used. Statistics reported for the summary effect sizes and associated heterogeneity are Cochran's $Q$ (a measure of heterogeneity in effect sizes) and its associated $p$-value, $I^2$ (a measure of the extent of heterogeneity, where up to 25% can be considered low; Higgins & Thompson, 2002), $\tau^2$ (an estimate of the between-study variance) and $H^2$ (a measure of the relative excess of $Q$ over its degrees of freedom which does not depend on the number of studies).

To deal with studies reporting multiple effect sizes from the same sample, but for different measures, a random data point was chosen for each of these studies for use in the overall analysis. This ensured that no participants were represented more than once in the overall analysis. Doing so avoids ignoring dependencies in the data that can bias estimates and cause the studies reporting multiple effect sizes to be given a heavier weighting in analysis. Where analyses based on test characteristics are undertaken, individual effect sizes from all specific tests are used because no study reported multiple results from the same test and sample (with the exception of studies 304A and 304B and 223A and 223B; study 304A was included when analyses were based on Figural Matrices and study 223A was included when analyses were based on Alphabetic Series).

Before conducting the analysis, moderator analysis was run to determine if there was significant heterogeneity in the effect sizes across Samples 1, 2, and 3. If the categorical variable "sample" was not found to account for significant heterogeneity in effect size, then this would indicate that the random samples were not systematically different and therefore should be representative of the broader body of literature. This would also indicate that results from these samples could be combined. A meta-regression model with sample as a categorical moderator variable was fit to the data. Sample was not a significant moderator of effect size ($Q(2) = 3.95$, $p = .139$, $R^2 = .02$). Sample as a moderator was also tested for stimuli and type categories, as well as individual tests, where numbers permitted. It was not a significant moderator of any of these, with the exception of

"Verbal Analogies" tests. However, this seemed to be due to an interaction between sample and test type, where Sample 2 contained results relevant to the Miller Analogies Test and Sample 3 contained results relevant to the Differential Aptitude Test –Verbal Reasoning. These individual tests showed very different effect sizes. Thus, given that no other significant moderator effects of sample were identified, samples were combined in all analyses.

## Results

### Descriptive Statistics

The final sample included data from 94 individual papers (see Supplementary Information for PRISMA diagrams) and 3 unpublished standardization or applicant data, totaling 115 separate samples (172 data points in total when accounting for studies reporting data for multiple tests). These samples provided cognitive ability data for 96,560 adults (51,035 males and 45,186 females; see Table 2 for more information). The samples ranged in size from 100 to 6,879 (*M* = 717.06, *SD* = 1033.63) when excluding the outlier of 14,815 for Study 401. Ratio of males to females included in studies was slightly biased towards a higher percentage of females (Percent Female: Range = 19% – 87%, *M* = 53%, *SD* = 15.4%). Study participants were mostly university students ($N_{studies}$ = 48), but a number were population representative samples ($N_{studies}$ = 18), community samples ($N_{studies}$ = 17), applicants to training programs or jobs ($N_{studies}$ = 9), school students ($N_{studies}$ = 9), or otherwise varied populations ($N_{studies}$ = 10; remainder unspecified or specific). Most studies reported results from English-speaking nations (UK, USA, Canada, Australia; $N_{studies}$ = 42), but a large proportion were European ($N_{studies}$ = 36), with the majority of these being Spanish. The remainder were from Africa, Asia and Latin America. Approximately half of the studies concerned young adult populations (ages 18-25); however, a substantial number included middle-aged to older adults ($N_{studies}$ = 47). A further $N_{studies}$ = 12 did not report the age of participants, but were included based on other information indicating they fell within the specified age range (e.g. described as university students). Publication year ranged from 1965 to 2016, with most being published within the last two decades ($M_{year}$ = 2006, $SD_{year}$ = 9.5). Date of data collection ranged from 1961 to 2015 ($M_{year}$ = 2001, $SD_{year}$ = 11.3).

Table 2

*Studies included in analysis*

| Authors | Year | ID | Percent Female | Age (Mean or Range) | *N* | Country | Population | Measure | Stimuli and Type | *g* |
|---|---|---|---|---|---|---|---|---|---|---|
| Abad et al. | 2004 | 301 | 46% | 17 - 30 | 1820 | Spain | Applicants | APM | FM | 0.27 |
| Abad et al.* | 2015 | 302A | 50% | 16-69 | 743 | Spain | Standardization | WAIS MR | FM | 0.09 |
| | | 302B | 50% | 16-69 | 743 | Spain | Standardization | WAIS S | VC | 0.06 |
| Abdel-Khalek & Lynn | 2009 | 201 | 40% | 18.0 | 220 | Saudi Arabia | School students | SPM | FM | 0.08 |
| | | 202 | 41% | 19.0 | 201 | Saudi Arabia | University | SPM | FM | 0.07 |
| | | 203 | 34% | 20-24 | 884 | Saudi Arabia | University | SPM | FM | -0.09 |
| Abdel-Khalek & Lynn | 2016 | 303 | 60% | 19.6 | 423 | Egypt | University | SPM | FM | 0.52 |
| Abdel-Khalek et al. | 2015 | 304A | 73% | 20.5 | 2147 | Egypt | University | APM | FM | 0.06 |
| Abdel-Khalek et al. | 2014 | 304B | 73% | 20.5 | 2147 | Egypt | University | SPM | FM | 0.06 |
| Ahmad et al. | 2008 | 101 | 51% | 31.7 | 2016 | Pakistan | Varied | SPM | FM | 0.04 |
| Al-Shahomee & Lynn | 2010 | 102 | 50% | 18-21 | 800 | Libya | University | SPM | FM | 0.00 |

| Authors | Year | ID | Percent<br>Female | Age (Mean<br>or Range) | *N* | Country | Population | Measure | Stimuli<br>and Type | *g* |
|---|---|---|---|---|---|---|---|---|---|---|
| Al-Shahomee & Lynn | 2012 | 305 | 50% | 38-50 | 520 | Libya | Community | SPM | FM | 0.35 |
| Alexopoulos | 1996 | 131 | 56% | 18 | 234 | Greece | School students | AH4 | V | 0.18 |
| Ali et al. | 2009 | 204 | 50% | 19.0 | 150 | Pakistan | University | SPM | FM | 0.40 |
| Bakhiet, Al-Qudah, et al. | 2016 | 307 | 37% | 18-25 | 1936 | Sudan | University | APM | FM | 0.08 |
| Bakhiet, Essa, et al. | 2016 | 308 | 50% | NR | 960 | Thailand | University | APM | FM | 0.08 |
| Bakhiet, Haseeb, et al. | 2015 | 309 | 42% | 18.0 | 347 | Sudan | School students | SPM | FM | 0.20 |
| Batey et al. | 2010 | 103 | 75% | 19.7 | 100 | UK | University | APM | FM | -0.39 |
| Birkett | 1980 | 205 | 58% | 22.5 | 103 | USA | Community | AH-4 | FV | 0.46 |
| Britton et al. | 2004 | 310 | 29% | 55.6 | 6033 | England | Community | AH-4 | V | 0.68 |
| Bromley | 1991 | 104 | 50% | 20-86 | 240 | England | Community | SPM | FM | 0.07 |
| Chamorro-Premuzic & Arteche | 2008 | 105 | 33% | 20.3 | 473 | UK | University | APM | FM | 0.06 |
| Chamorro-Premuzic et al. | 2005 | 206 | 73% | 19.6 | 181 | UK & USA | University | SPM | FM | 0.00 |
| Chamorro-Premuzic et al. | 2009 | 207 | 81% | 20.1 | 248 | Spain | University | Gf | NA | -0.08 |

| Authors | Year | ID | Percent Female | Age (Mean or Range) | *N* | Country | Population | Measure | Stimuli and Type | *g* |
|---|---|---|---|---|---|---|---|---|---|---|
| Choi & L'Hirondelle | 2005 | 208 | 55% | 19.3 | 111 | Canada | University | APM | FM | 0.38 |
| Colom et al.* | 2008 | 314A | 71% | 18.0 | 111 | Spain | School & University | PMA-R | AS | -0.28 |
|  |  | 314B | 71% | 18.0 | 111 | Spain | School & University | DAT-AR | FS | 0.12 |
|  |  | 314C | 71% | 18.0 | 111 | Spain | School & University | DAT-VR | VA | -0.54 |
|  |  | 315A | 83% | 20.2 | 261 | Spain | University | PMA-R | AS | -0.02 |
|  |  | 315B | 83% | 20.2 | 261 | Spain | University | DAT-AR | FS | 0.51 |
|  |  | 315C | 83% | 20.2 | 261 | Spain | University | DAT-VR | VA | 0.47 |
|  |  | 316A | 81% | 20.3 | 289 | Spain | University | PMA-R | AS | 0.20 |
|  |  | 316B | 81% | 20.3 | 289 | Spain | University | DAT-AR | FS | 0.58 |
|  |  | 316C | 81% | 20.3 | 289 | Spain | University | DAT-VR | VA | 0.59 |

| Authors | Year | ID | Percent Female | Age (Mean or Range) | *N* | Country | Population | Measure | Stimuli and Type | *g* |
|---|---|---|---|---|---|---|---|---|---|---|
| Colom et al.* | 2002 | 317 | 86% | 18.8 | 104 | Spain | University | CFIT | FV | 0.23 |
| Colom et al. | 2004 | 318 | 50% | 19.9 | 239 | Spain | University | APM | FM | 0.29 |
| Colom & Garcia-Lopez | 2002 | 311A | 50% | 18.4 | 604 | Spain | Applicants | APM | FM | 0.30 |
| | | 311B | 50% | 18.4 | 604 | Spain | Applicants | CFIT | FV | 0.10 |
| Colom et al. | 2000 | 319A | 42% | 23.1 | 3596 | Spain | Applicants | PMA-R | AS | -0.19 |
| | | 319B | 42% | 23.1 | 3596 | Spain | Applicants | DAT-VR | VA | 0.22 |
| | | 320 | 40% | 23.1 | 6879 | Spain | Applicants | PMA-R | AS | -0.11 |
| Colom & Lynn | 2004 | 312A | 52% | 18.0 | 151 | Spain | Standardization | DAT-VR | VA | 0.37 |
| | | 312B | 53% | 18.0 | 149 | Spain | Standardization | DAT-AR | FS | 0.36 |
| Colom et al.* | 2015 | 321 | 77% | 19.0 | 302 | Spain | University | DAT-AR | FS | 0.50 |
| Colom & Quiroga* | 2009 | 313A | 80% | 19.9 | 198 | Spain | University | APM | FM | 0.16 |
| | | 313B | 80% | 19.9 | 198 | Spain | University | PMA-R | AS | -0.34 |
| | | 313C | 80% | 19.9 | 198 | Spain | University | DAT-AR | FS | 0.19 |

| Authors | Year | ID | Percent Female | Age (Mean or Range) | *N* | Country | Population | Measure | Stimuli and Type | *g* |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 313D | 80% | 19.9 | 198 | Spain | University | DAT-VR | VA | 0.26 |
| Colom et al. | 1999 | 322A | 30% | 18.0 | 1417 | Spain | School students | DAT-VR | VA | 0.30 |
| | | 322B | 31% | 18.0 | 1416 | Spain | School students | DAT-AR | FS | 0.67 |
| | | 323A | 41% | 18.0 | 2726 | Spain | School students | DAT-VR | VA | 0.31 |
| | | 323B | 40% | 18.0 | 2711 | Spain | School students | DAT-AR | FS | 0.40 |
| | | 324 | 37% | 18.0 | 180 | Spain | School students | PMA-R | AS | -0.38 |
| | | 325 | 20% | 18.0 | 701 | Spain | School students | PMA-R | AS | -0.35 |
| Colom et al. | 2013 | 326A | 57% | 19.9 | 104 | Spain | University | Gf | NA | -0.21 |
| | | 326B | 57% | 19.9 | 104 | Spain | University | APM | FM | -0.06 |
| | | 326C | 57% | 19.9 | 104 | Spain | University | PMA-R | AS | -0.49 |
| | | 326D | 57% | 19.9 | 104 | Spain | University | DAT-AR | FS | 0.17 |
| | | 326E | 57% | 19.9 | 104 | Spain | University | DAT-VR | VA | 0.17 |
| Cvorovic & Lynn | 2014 | 327 | 54% | 54.5 | 136 | Serbia | Community | SPM | FM | 0.27 |

| Authors | Year | ID | Percent Female | Age (Mean or Range) | *N* | Country | Population | Measure | Stimuli and Type | *g* |
|---|---|---|---|---|---|---|---|---|---|---|
| Deary et al. | 2001 | 328 | 54% | 55.0 | 900 | Scotland | Population | AH-4 | V | 0.09 |
| Der et al. | 2012 | 329 | 54% | 57.0 | 825 | Scotland | Population | AH-4 | V | 0.12 |
| | | 330 | 53% | 36.6 | 757 | Scotland | Population | AH-4 | V | 0.14 |
| Diaz & Lynn | 2016 | 331A | 52% | 16-90 | 887 | Chile | Standardization | WAIS MR | FM | 0.15 |
| | | 331B | 52% | 16-90 | 887 | Chile | Standardization | WAIS S | VC | 0.01 |
| Diaz et al. | 2010 | 106 | 61% | 25.2 | 258 | Spain | Varied | SPM | FM | 0.19 |
| Dolan et al. | 2006 | 332A | 48% | 16-34 | 588 | Spain | Standardization | WAIS MR | FM | 0.12 |
| | | 332B | 48% | 16-34 | 588 | Spain | Standardization | WAIS S | VC | -0.02 |
| Escorial et al. | 2003 | 333A | 51% | 33.1 | 403 | Spain | Standardization | WAIS S | VC | 0.11 |
| | | 333B | 51% | 33.1 | 403 | Spain | Standardization | WAIS MR | FM | 0.33 |
| Essa et al. | 2016 | 334 | 57% | 20.0 | 1502 | Egypt | University | APM | FM | 0.23 |
| Estrada et al.* | 2015 | 335A | 82% | 20.1 | 477 | Spain | University | APM | FM | 0.24 |
| | | 335B | 82% | 20.1 | 477 | Spain | University | DAT-AR | FS | 0.16 |

| Authors | Year | ID | Percent Female | Age (Mean or Range) | *N* | Country | Population | Measure | Stimuli and Type | *g* |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 335C | 82% | 20.1 | 477 | Spain | University | DAT-VR | VA | 0.24 |
| Flores-Mendoza et al. | 2016 | 336 | 63% | 22.3 | 1042 | Brazil | University | APM | FM | 0.43 |
| Flynn & Rossi-Case | 2011 | 337 | 51% | 19-20 | 229 | Argentina | University | SPM | FM | 0.02 |
| | | 338 | 50% | 21-22 | 242 | Argentina | University | SPM | FM | 0.01 |
| | | 339 | 51% | 23-24 | 217 | Argentina | University | SPM | FM | -0.02 |
| | | 340 | 51% | 25-30 | 305 | Argentina | University | SPM | FM | -0.01 |
| Foley & Proff | 1965 | 209 | 28% | NA | 462 | USA | Trainees | MAT | VA | -0.07 |
| | | 210 | 22% | NA | 447 | USA | Trainees | MAT | VA | -0.06 |
| Furnham et al. | 2008 | 107 | 32% | 36.5 | 188 | UK | Applicants | SPM | FM | -0.10 |
| Gale et al. | 2011 | 341 | 51% | 64.4 | 418 | Scotland | Community | SPM | FM | 0.16 |
| Gangestad et al. | 2010 | 108 | 50% | 21.5 | 132 | USA | Varied | APM | FM | 0.05 |
| GeneSys Australia* | 2015 | 401A | 36% | 18-64 | 14815 | Australia | Applicants | GRT2 AR | FV | 0.22 |
| | | 401B | 39% | 18-64 | 14815 | Australia | Applicants | GRT2 NR | NV | 0.57 |

| Authors | Year | ID | Percent Female | Age (Mean or Range) | *N* | Country | Population | Measure | Stimuli and Type | *g* |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 401C | 36% | 18-64 | 14815 | Australia | Applicants | GRT2 VR | VV | 0.02 |
| Hakstian & Cattell | 1982 | 402 | 55% | NA | 190 | USA | University | CAB-I | AC | 0.04 |
| Hambrick et al. | 2010 | 109A | 71% | NA | 109 | US | University | APM | FM | 0.04 |
| | | 109B | 71% | NA | 109 | USA | University | Letter Sets (Ekstrom) | AC | -0.29 |
| Hambrick et al. | 2008 | 110A | 72% | NA | 516 | USA | University | SILS Abstraction | AV | 0.18 |
| | | 110B | 72% | NA | 516 | USA | University | Letter Sets (Ekstrom) | AC | -0.07 |
| | | 110C | 72% | NA | 516 | USA | University | APM | FM | 0.35 |
| Hattori & Lynn | 1997 | 342 | 50% | 16-74 | 1402 | Japan | Standardization | WAIS S | VC | 0.09 |
| Hegarty et al. | 2009 | 211 | 39% | NA | 206 | USA | University | DAT-AR | FS | 0.18 |
| House & Keeley | 1995 | 212 | 87% | NA | 1438 | USA | University | MAT | VA | 0.17 |

| Authors | Year | ID | Percent Female | Age (Mean or Range) | *N* | Country | Population | Measure | Stimuli and Type | *g* |
|---|---|---|---|---|---|---|---|---|---|---|
| Johnson & Bouchard* | 2007 | 343A | 58% | 42.7 | 427 | North America, Great Britain and Australia | Varied | CAB-I | AC | 0.01 |
| | | 343B | 57% | 42.7 | 395 | | Varied | WAIS S | VC | 0.02 |
| | | 343C | 61% | 42.7 | 299 | | Varied | RPM | FM | 0.04 |
| Kagan & Stock | 1980 | 213 | 44% | NA | 154 | USA | Applicants | MAT | VA | 0.16 |
| Kaufman & Horn | 1996 | 111A | 52% | 44.5 | 1472 | USA | Standardization | KAIT Logical Steps | NA | 0.19 |
| | | 111B | 52% | 44.5 | 1472 | USA | Standardization | KAIT Mystery Codes | NA | 0.13 |
| | | 111C | 52% | 44.5 | 1500 | USA | Standardization | KAIT Fluid IQ | NA | 0.15 |
| Kaufman & Kaufman | 2004 | 403 | 50% | 18-90 | 620 | USA | Standardization | KBIT Matrices | FM | 0.08 |
| Keith et al.* | 2008 | 344A | 57% | 18-59 | 2476 | USA | Standardization | WJIII Concept Formation | NA | -0.17 |

| Authors | Year | ID | Percent Female | Age (Mean or Range) | *N* | Country | Population | Measure | Stimuli and Type | *g* |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 344B | 57% | 18-59 | 2024 | USA | Standardization | WJIII Number Series | NS | 0.07 |
| | | 344C | 60% | 18-59 | 1896 | USA | Standardization | WJIII Number Matrices | NM | 0.22 |
| Khaleefa et al. | 2010 | 113 | 52% | NA | 240 | Sudan | Varied | SPM | FM | 0.12 |
| Khaleefa et al. | 2014 | 345 | 70% | 16-19 | 1001 | Sudan | University | SPM | FM | 0.22 |
| Khaleefa et al. | 2008 | 114 | 45% | 18.0 | 310 | Sudan | School students | SPM | FM | -0.25 |
| | | 115 | 65% | 20-25 | 2984 | Sudan | University | SPM | FM | 0.09 |
| | | 116 | 55% | 19.0 | 408 | Sudan | University | SPM | FM | 0.19 |
| Khaleefa & Lynn | 2008 | 112 | 30% | 18.0 | 110 | Syria | NA | SPM | FM | 0.24 |
| Lynn | 1998 | 347 | 52% | 44.3 | 200 | Scotland | Standardization | WAIS S | VC | 0.23 |
| Lynn | 2014 | 346 | 45% | 19.0 | 310 | Cambodia | University | SPM | FM | 0.07 |

| Authors | Year | ID | Percent Female | Age (Mean or Range) | *N* | Country | Population | Measure | Stimuli and Type | *g* |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 360 | 61% | 18.5 | 346 | Germany | School & University | SPM | FM | 0.27 |
| Lynn & Dai | 1993 | 348 | 42% | 16-65+ | 1979 | China | Standardization | WAIS S | VC | 0.15 |
| Lynn & Hur | 2016 | 349A | 50% | 20-34 | 424 | South Korea | Standardization | WAIS MR | FM | 0.49 |
| | | 349B | 50% | 20-34 | 424 | South Korea | Standardization | WAIS S | VC | 0.41 |
| | | 350A | 57% | 35-69 | 486 | South Korea | Standardization | WAIS MR | FM | 0.01 |
| | | 350B | 57% | 35-69 | 486 | South Korea | Standardization | WAIS S | VC | 0.17 |
| Lynn & Irwing | 2004 | 351 | 19% | 20.0 | 2222 | USA | University | APM | FM | 0.23 |
| Lynn & Tse-Chan | 2003 | 352 | 29% | 18.0 | 196 | Hong Kong | School students | APM | FM | 0.38 |
| MacCann | 2010 | 117A | 76% | 21.9 | 147 | Australia | University | Esoteric Analogies | VA | 0.45 |
| | | 117B | 76% | 21.9 | 147 | Australia | University | Letter Series | AS | 0.06 |
| Martinez et al.* | 2011 | 355A | 80% | 19.7 | 182 | Spain | University | APM | FM | 0.17 |

| Authors | Year | ID | Percent Female | Age (Mean or Range) | *N* | Country | Population | Measure | Stimuli and Type | *g* |
|---|---|---|---|---|---|---|---|---|---|---|
|  |  | 355B | 80% | 19.7 | 182 | Spain | University | PMA-R | AS | -0.18 |
|  |  | 355C | 80% | 19.7 | 182 | Spain | University | DAT-AR | FS | 0.34 |
|  |  | 355D | 80% | 19.7 | 182 | Spain | University | DAT-VR | VA | 0.44 |
| Martinez & Colom* | 2009 | 354A | 82% | 20.1 | 265 | Spain | University | APM | FM | 0.42 |
|  |  | 354B | 82% | 20.1 | 265 | Spain | University | PMA-R | AS | -0.15 |
|  |  | 354C | 82% | 20.1 | 265 | Spain | University | DAT-AR | FS | 0.56 |
|  |  | 354D | 82% | 20.1 | 265 | Spain | University | DAT-VR | VA | 0.52 |
| Naderi et al. | 2010 | 119 | 31% | 18-27 | 153 | Iranian students in Malaysia | University | CFIT | FV | 0.13 |
| Parker et al. | 1991 | 120 | 31% | 18+ | 683 | USA | Community | SILS Abstraction | AV | -0.23 |
| Piffer | 2016 | 356A | 52% | 16-90 | 2200 | USA | Standardization | WAIS MR | FM | 0.07 |
|  |  | 356B | 52% | 16-90 | 2200 | USA | Standardization | WAIS S | VC | 0.11 |

| Authors | Year | ID | Percent Female | Age (Mean or Range) | *N* | Country | Population | Measure | Stimuli and Type | *g* |
|---|---|---|---|---|---|---|---|---|---|---|
| Ponton et al. | 1996 | 121 | 61% | 16-75 | 171 | USA | Community | SPM | FM | 0.21 |
| Quiroga et al.* | 2015 | 306A | 64% | 22.2 | 185 | Spain | University | APM | FM | 0.32 |
| | | 306B | 64% | 22.2 | 185 | Spain | University | DAT-AR | FS | 0.14 |
| | | 306C | 64% | 22.2 | 185 | Spain | University | DAT-VR | VA | 0.35 |
| Rammsayer & Troche | 2010 | 122 | 52% | 25.1 | 276 | Switzerland | Varied | CFIT | FV | 0.14 |
| Raz et al. | 2009 | 214 | 66% | 54.4 | 189 | USA | Community | CFIT | FV | 0.36 |
| Rushton & Čvorović | 2009 | 215 | 32% | 17-60 | 130 | Serbia – Novi Pazar | Community | SPM | FM | 0.08 |
| | | 216 | 28% | 17-50 | 404 | Serbia - Belgrade | Community | SPM | FM | -0.13 |
| Saccuzzo et al. | 1996 | 123 | 64% | NA | 240 | USA | University | APM | FM | -0.18 |
| Salthouse & Mitchell | 1990 | 217A | 51% | 20-83 | 383 | USA | Community | SILS Abstraction | AV | 0.14 |

| Authors | Year | ID | Percent Female | Age (Mean or Range) | *N* | Country | Population | Measure | Stimuli and Type | *g* |
|---|---|---|---|---|---|---|---|---|---|---|
|  |  | 217B | 51% | 20-83 | 383 | USA | Community | Letter Sets (Ekstrom) | AC | -0.02 |
| Schaie et al. | 2005 | 223A | 59% | 64.2 | 180 | USA | Community | PMA-R | AS | -0.11 |
|  |  | 223B | 59% | 64.2 | 180 | USA | Community | Letter Series | AS | 0.15 |
|  |  | 223C | 59% | 64.2 | 180 | USA | Community | Word Series | WS | -0.05 |
|  |  | 223D | 59% | 64.2 | 180 | USA | Community | Number Series | NS | 0.63 |
| Sellami et al. | 2010 | 124 | 42% | 26.6 | 115 | Morocco | University | SPM | FM | 0.05 |
| Silvia & Sanders | 2010 | 125A | 74% | NA | 129 | USA | University | APM | FM | 0.26 |
|  |  | 125B | 74% | NA | 129 | USA | University | Letter Sets (Ekstrom) | AC | 0.24 |
| Stein et al. | 2005 | 218 | 56% | 35-41 | 521 | Guatemala | Specific | SPM | FM | 0.46 |
|  |  | 219 | 52% | 26-34 | 948 | Guatemala | Specific | SPM | FM | 0.56 |
| Stewart et al. | 2006 | 357 | 73% | 64.3 | 504 | Scotland | Community | SPM | FM | 0.27 |

| Authors | Year | ID | Percent Female | Age (Mean or Range) | *N* | Country | Population | Measure | Stimuli and Type | *g* |
|---|---|---|---|---|---|---|---|---|---|---|
| Tan | 1991 | 126 | 37% | 18 - 21 | 197 | NR | NA | CFIT | FV | 0.08 |
| Tommasi et al. * | 2015 | 358 | 50% | 16-64 | 1630 | Italy | Standardization | WAIS S | VC | 0.01 |
| Tuttle & Pillard | 1991 | 127 | 41% | 25-40 | 164 | USA | Community | WAIS S | VC | 0.05 |
| van der Sluis et al. | 2006 | 359A | 56% | 18-46 | 517 | Netherlands | Varied | WAIS S | VC | 0.18 |
| | | 359B | 56% | 18-46 | 522 | Netherlands | Varied | WAIS MR | FM | 0.20 |
| van Leeuwen et al. | 2008 | 128 | 50% | 42.8 | 189 | Netherlands | Varied | APM | FM | 0.18 |
| Vigneau & Bors | 2005 | 220 | 65% | 19.9 | 644 | Canada | University | APM | FM | 0.17 |
| Vigneau & Bors | 2008 | 129 | 64% | 20.0 | 506 | Canada | University | APM | FM | 0.25 |
| von Stumm et al. | 2009 | 221A | 81% | 20.1 | 246 | Spain | University | APM | FM | 0.02 |
| | | 221B | 81% | 20.1 | 246 | Spain | University | PMA-R | AS | -0.34 |
| | | 221C | 81% | 20.1 | 246 | Spain | University | DAT-AR | FS | 0.06 |
| | | 221D | 81% | 20.1 | 246 | Spain | University | DAT-VR | VA | 0.31 |
| Wachs et al. | 1996 | 130 | 66% | 32.2 | 152 | Egypt | Community | SPM | FM | 0.48 |

| Authors | Year | ID | Percent Female | Age (Mean or Range) | *N* | Country | Population | Measure | Stimuli and Type | *g* |
|---|---|---|---|---|---|---|---|---|---|---|
| Welborn et al. | 2009 | 222 | 50% | 22.8 | 117 | USA | Community | APM | FM | 0.12 |

*Note.* APM = Advanced Progressive Matrices; AH-4 = Alice Heim Test; CAB-I = Comprehensive Ability Battery – Inductive Reasoning; CFIT = Culture Fair Intelligence Test; DAT-AR = Differential Aptitude Test – Abstract Reasoning; DAT-VR: Differential Aptitude Test – Verbal Reasoning; Gf = latent Gf factor; GRT2 = General Reasoning Test (AR = Abstract Reasoning; NR = Numerical Reasoning; VR = Verbal Reasoning); KAIT = Kaufmann Adult Intelligence Test; KBIT = Kaufmann Brief Intelligence Test; MAT = Miller Analogies Test; PMA-R = Primary Mental Abilities – Reasoning; RPM = Raven's Progressive Matrices (Advanced or Standard not specified); SILS = Shipley Institute of Living Scale; SPM = Standard Progressive Matrices; WAIS-MR = Wechsler Adult Intelligence Scale Matrix Reasoning; WAIS-S = Wechsler Adult Intelligence Scale Similarities. FV = Figural various; FM = Figural matrices; AS = Alphabetic Series; FS = Figural series; VA = Verbal analogies; VS = Verbal Similarities; V = Various. NA = Not applicable or information not reported. * = Data obtained from personal communication with author

**Heterogeneity and Subgroup Analyses**

Significant heterogeneity was found between the mean effect sizes obtained from individual studies ($g$ = 0.13; $Q$(114) = 1010.05, $p$ < .001, $I^2$ = 84.26). It was noted that the PMA-R demonstrated an effect size opposite to the large majority of the data; therefore heterogeneity excluding the PMA-R was also assessed. Significant heterogeneity was still present ($Q$(108) = 727.06, $p$ < .001, $I^2$ = 80.09). Given that the current study was interested in investigating the role of stimulus and question type on the size of the sex difference, these categorical variables were examined as moderators.

Stimuli considered were alphabetic, figural and verbal, and types considered were analogies, classification, matrices and series (see Table 3). Stimulus ($Q_\mathrm{M}$(2) = 28.88, $p$ < .001, $R^2$ = .32) was found to be a significant moderator while type approached nominal significance ($Q_\mathrm{M}$(3) = 7.02, $p$ = .071, $R^2$ = .07). When accounting for both stimulus and type, figural stimulus was found to significantly favor males as compared to alphabetic (0.51 decrease in effect size, $p$ < .001) and verbal (0.41 decrease in effect size, $p$ = .003). Results indicated that significant heterogeneity in effect sizes remained after accounting for stimulus and type ($Q_\mathrm{E}$(90) = 263.70, $p$ < .001, $I^2$ = 69.04).

Irwing and Lynn (2004b) report a significant moderation effect of the relative “selectivity” of males and females, operationalized as the difference in the variability of their scores. Here, higher male variability versus higher female variability was not a significant overall moderator ($Q_\mathrm{M}$(1) = 0.17, $p$ = .677, $b$ = -0.02). However, proportion of females in the sample was nominally significant as a moderator of the effect size, with an increase in effect size favoring males as a function of greater proportion of females ($Q_\mathrm{M}$(1) = 4.04, $p$ = .044, $b$ = 0.26; see sensitivity analysis).

Additional moderators considered were participant education (university, varied, high-school, pre-high school), type of population (university students and other), whether or not the study aimed to investigate sex differences, participant age, variability of male versus female groups, data collection date and country (categorized according to continent). Only date of data collection was found to slightly increase effect size ($Q_M(1) = 8.58$, $p = .003$, $b = 0.01$). Test categories and individual tests are considered in the following section.

Table 3

*Number of data points for each test category*

| | Alphabetic | Figural | Verbal | Various | Total |
|---|---|---|---|---|---|
| Analogies | | | 8 *(19)* | | 8 *(19)* |
| Classification | 5 *(6)* | | 9 *(14)* | | 14 *(20)* |
| Matrices | | 59 *(75)* | | | 59 *(75)* |
| Series | 7 *(15)* | 8 *(15)* | | 2 *(3)* | 17 *(33)* |
| Various | | 8 *(8)* | *(1)* | 5 *(5)* | 13 *(14)* |
| Total | 12 *(21)* | 75 *(98)* | 17 *(34)* | 7 *(8)* | |

*Note:* Numerical stimulus category not included. Values outside parentheses indicate the number of data points included in the overall analysis; values in parentheses indicate total number of data points used in test category analysis.

**Test Categories.** Based on the categorizations in Table 3, six test classifications were produced with sufficient numbers for analysis: Figural Matrices, Figural (Various), Figural Series, Verbal Analogies, Verbal Classification and Alphabetic Series. Table 4 presents the summary effect sizes and heterogeneity statistics for these groupings, while Figures 1 to 8 present the associated forest plots. Results are presented for all studies using each measure, and where applicable, for university samples only. All but Figural (Various) demonstrated significant heterogeneity. Figural (Various) tests showed a weak-to-moderate male advantage, although the number of studies in this group was small. Further investigation of the other test classification groups was undertaken by examining tests at the individual test level when number of data points allowed (for example, figural matrices tests were divided into APM, SPM and WAIS Matrix Reasoning groups). This was not undertaken with Verbal Classification, because the only verbal classification test included in the present study was the WAIS Similarities test. In many cases, although examining results at the test level did not account for all heterogeneity, the amount of heterogeneity as assessed by the $I^2$ value was decreased.

Table 4

*Summary statistics by test category and test*

| Category | Hedge's *g* | *Q (df)* | *p* | $H^2$ | $\tau^2$ | $I^2$ |
|---|---|---|---|---|---|---|
| Figural Matrices | 0.16 [.12, .20] | 230.29 *(74)* | < .001 | 3.30 | 0.02 | 69.7 |
| *APM* | 0.19 [.13, .24] | 63.91 *(28)* | < .001 | 2.25 | 0.01 | 55.6 |
| *APM (University students)* | 0.17 [.11, .24] | 55.84 *(22)* | < .001 | 2.54 | 0.01 | 60.7 |
| *APM (Other populations)* | 0.26 [.19, .33] | 3.22 *(5)* | .667 | 1 | < .01 | 0 |
| *SPM*[1] | 0.11 [.06, .17] | 76.40 *(33)* | < .001 | 2.66 | 0.01 | 62.3 |
| *SPM (University students)* | 0.10 [.02, .17] | 36.87 *(15)* | .001 | 2.92 | .01 | 65.7 |
| *SPM (Other populations)*[1] | 0.13 [.05, .22] | 38.80 *(17)* | .002 | 2.36 | 0.02 | 57.7 |
| *WAIS Matrix Reasoning* | 0.17 [.07, .27] | 21.98 *(7)* | .003 | 3.56 | 0.02 | 71.9 |
| Figural (Various) | 0.21 [.16, .26] | 6.07 *(7)* | .532 | 1.06 | < .01 | 5.7 |
| *CFIT* | 0.14 [.04, .25] | 2.55 *(5)* | .769 | 1 | < .01 | 0 |
| Figural Series | 0.35 [.24, .46] | 42.54 *(14)* | < .001 | 2.98 | 0.03 | 66.4 |
| *DAT-AR*[2] | 0.32 [.23, .41] | 19.70 *(13)* | .103 | 1.56 | 0.01 | 35.9 |
| *DAT-AR (University students)* | 0.31 [.19, .43] | 16.04 *(10)* | .098 | 1.63 | 0.01 | 38.6 |

| Category | Hedge's *g* | *Q (df)* | *p* | $H^2$ | $\tau^2$ | $I^2$ |
|---|---|---|---|---|---|---|
| Verbal Analogies | 0.24 [.15, .34] | 49.22 *(18)* | < .001 | 4.05 | 0.03 | 75.3 |
| *DAT-VR* | 0.29 [.23, .36] | 28.28 *(13)* | .008 | 1.54 | < .01 | 34.9 |
| *DAT-VR (University students)* | 0.37 [.27, .47] | 6.25 *(8)* | .619 | 1 | < .01 | 0 |
| Verbal Classification[3] | 0.10 [.05, .15] | 21.74 *(13)* | .060 | 1.61 | < .01 | 37.8 |
| Alphabetic Series | -0.18 [-.26, -.10] | 23.89 *(13)* | .032 | 2.05 | 0.01 | 51.2 |
| *PMA-R* | -0.19 [-.27, -.11] | 22.70 *(12)* | .030 | 2.15 | .01 | 53.6 |
| *PMA-R (University students)* | -0.17 [-.35, .00] | 11.61 *(6)* | .071 | 1.93 | .03 | 48.3 |

*Note*. Positive Hedges *g* signifies male advantage. $H^2$ indicates the relative excess of *Q* over its *df*; $I^2$ indicates the percentage of variation between studies that is due to significant heterogeneity; $\tau^2$ indicates the between-study variance of the effects, where the square root of this value is the underlying standard deviation (in the Hedges *g* metric) across studies. [1]Excludes Brazilian data; [2]Excludes 1979 Spanish standardization; [3]WAIS Similarities only; APM = Advanced Progressive Matrices; DAT-AR = Differential Aptitude Test – Abstract Reasoning; DAT-VR: Differential Aptitude Test – Verbal Reasoning; PMA-R = Primary Mental Abilities – Reasoning; SPM = Standard Progressive Matrices.

***APM.*** Effect sizes for the APM ranged from a moderate female advantage to a moderate male advantage (-0.39 to 0.43). Unsurprisingly, significant heterogeneity was identified in the effect sizes for this test (see Table 4). Only three out of the 29 studies concerning the APM reported an effect size favoring females (Studies 103: *g* = -0.39; 123: *g* = -0.18; and 326B: *g* = -0.06). It was unclear, however, what characteristics of these studies may have caused this. They did not differ systematically from the others in the age range, education, sampling procedure, country, population or administration characteristics (number of items, timed versus untimed). Difference in variability of males and females was not a significant moderator of the sex effect on APM ($Q_M(1) = 0.06$, $p = .811$, $b = -0.02$, $R^2 = .00$). Country, however, was a significant moderator ($Q_M(4) = 10.29$, $p = .036$, $R^2 = .50$), accounting for much of the residual heterogeneity ($Q_E$ $(24) = 38.46$, $p = .031$). This was due to a particularly large effect size found in the Brazilian sample ($g = 0.43$). Once this study was excluded, country was no longer a significant moderator of the overall APM effect size, although significant heterogeneity remained. No other variables considered were found to account for this heterogeneity. When APM effect sizes were considered in non-university samples only, significant heterogeneity was not identified, but this may be due to the low number of data points (only 6 studies reported results for the APM not using university samples).

Results for the APM are presented for university samples to enable comparison between the current results and those of Irwing and Lynn (2005). As can be seen in Table 4, within the university population the effect sizes of the sex differences remained heterogeneous, and furthermore, the summary effect was somewhat lower than that reported by Irwing and Lynn. It did, however, favor males. The effect size was larger in populations with greater male variability (*g* = 0.18 versus 0.11), but variability as a moderator was found to be non-significant ($Q_M(1) = 1.15$, $p = .284$, $b = 0.07$, $R^2 = .00$).

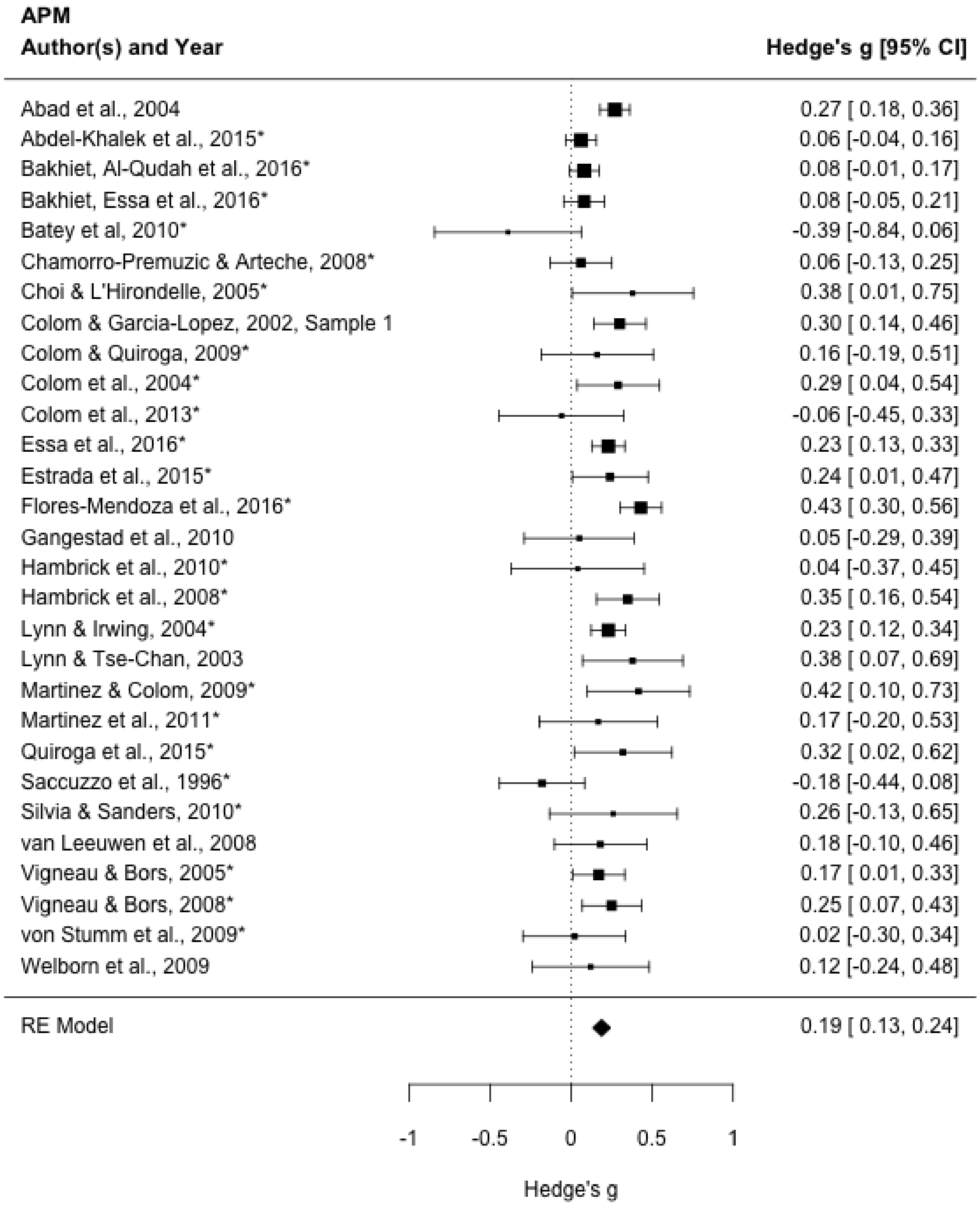


*Figure 1.* Forest plot of effect sizes for the Advanced Progressive Matrices. * indicates university student sample.

***SPM***. The summary effect for the SPM demonstrated a small male advantage, but significant heterogeneity was identified. Date of data collection was found to be a significant moderator of the effect size in university populations, with effect size increasing as a function of increasing year ($Q_{\mathrm{M}}(1) = 6.89$, $p = .009$, $b = 0.01$, $R^2 = .57$).

An examination of the effect sizes for the SPM in non-university populations initially indicated that country was a significant moderator of the effect size. However, on closer inspection it appeared that the significant moderation effect occurred because of one South American sample (Study 218 & 219) that showed a particularly large male advantage. This study was conducted using participants from a nutritional supplementation program in rural Guatemala. In this sample, schooling attainment was much lower for women, and increases in school attainment were related to increases in RPM scores indicating a disadvantaged situation for the women, which may have contributed to the particularly large sex difference. Once this study was removed, there were no differences in effect size across country, however significant heterogeneity was still present in the SPM effect size (see Table 4). The summary effect size for the SPM in non-university populations was slightly larger than that for university students.

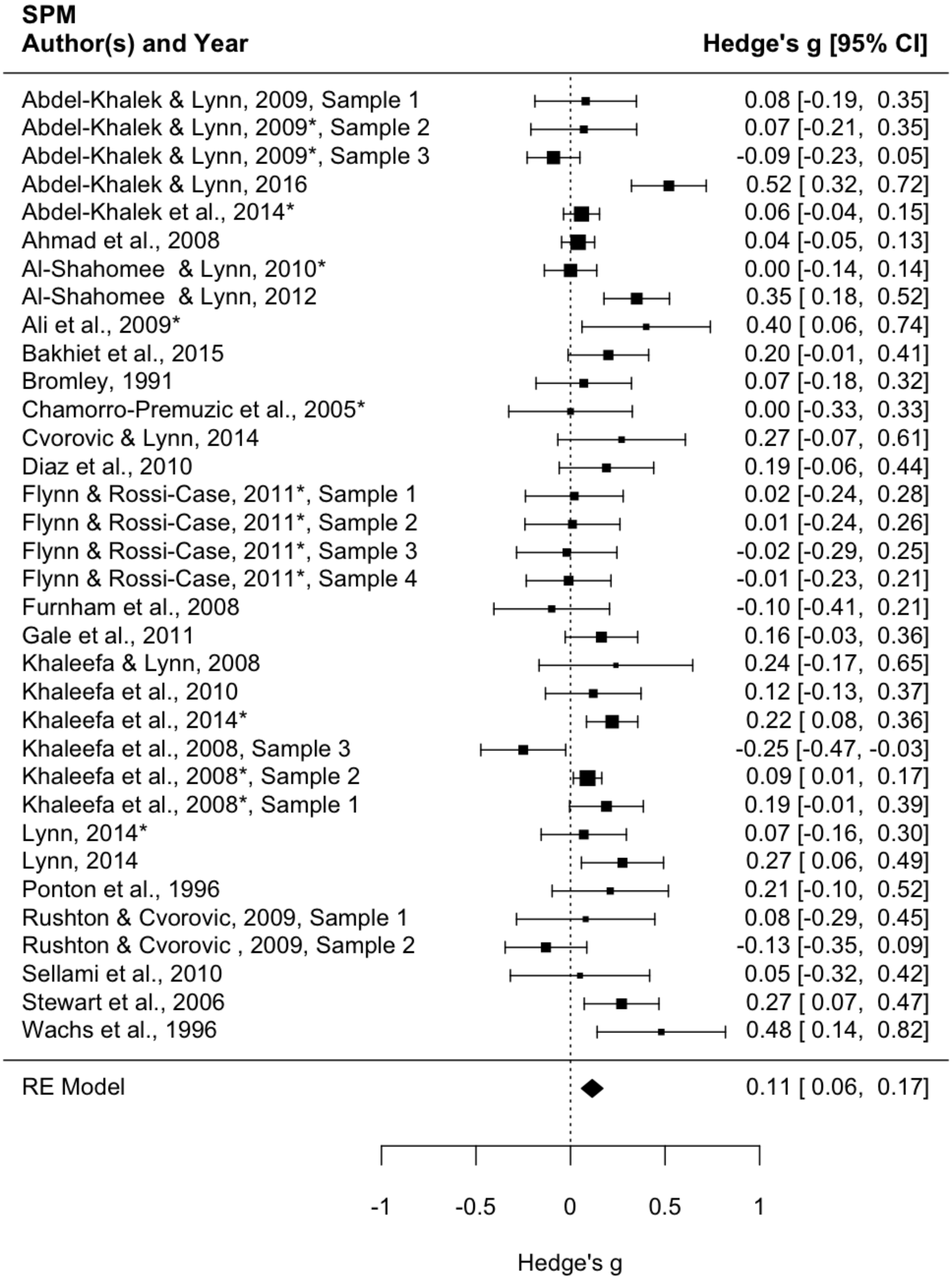


*Figure 2.* Forest plot of effect sizes for the Standard Progressive Matrices. * indicates university student sample.

***WAIS Matrix Reasoning.*** Effect sizes for the WAIS Matrix Reasoning subtest indicated a small-to-moderate male advantage, although again significant heterogeneity was identified in the effect sizes reported. This is particularly interesting in this case, because the majority of these samples were representative standardization samples, and it is thus difficult to make the claim that this heterogeneity is due to the use of unrepresentative or biased samples. Furthermore, the effect size most strongly favoring females, and the effect size most strongly favoring males, were both from the South Korean standardization, albeit different age groups (see Figure 3). Although it is difficult to tell with a limited range of data points and limited specificity regarding age, it may be that the effect size varies with age. Three of the four samples classified as young-to-middle aged adults demonstrated the strongest male advantage. None of the variables considered as moderators were found to account for the heterogeneity in effect size.

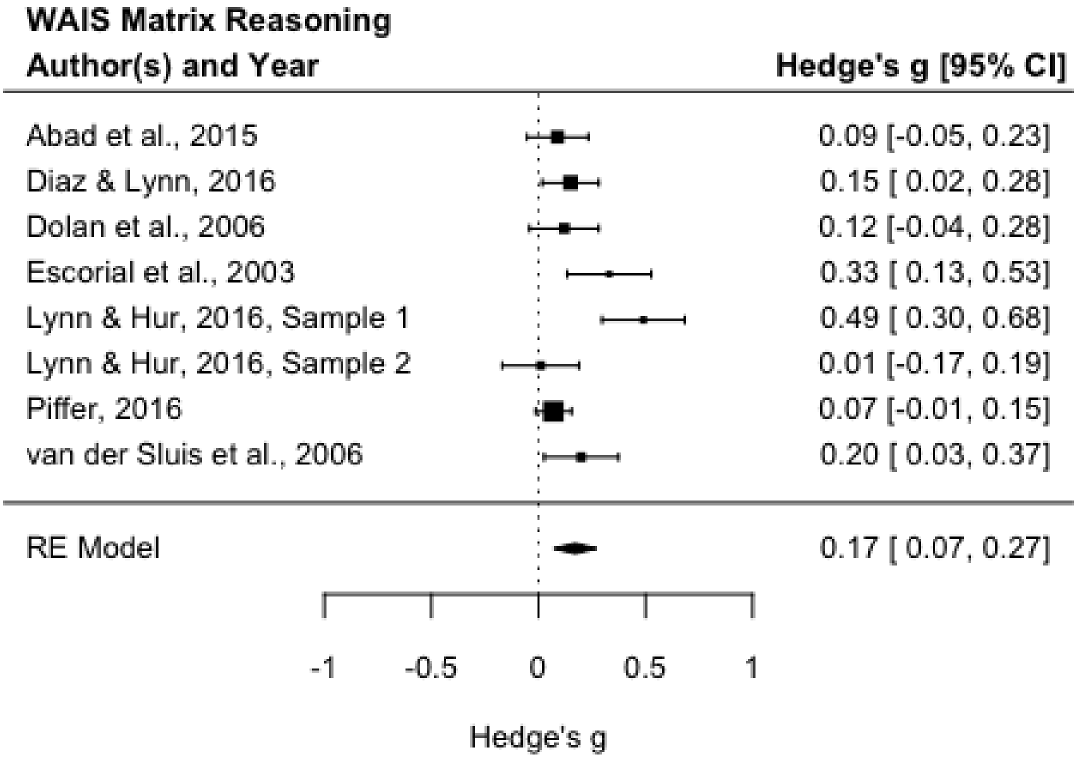


*Figure 3.* Forest plot of effect sizes for the WAIS – Matrix Reasoning.

***DAT-AR.*** The summary effect size for the DAT-AR showed the strongest male advantage of all individual measures considered, although, again, significant heterogeneity was present in effect sizes. Year of data collection was found to be a significant moderator of the effect size, accounting for all residual heterogeneity ($Q_{\mathrm{M}}(1) = 10.23$, $p = .001$, $R^2 = .73$; $Q_{\mathrm{E}}(13) = 18.92$, $p = .126$). There was a slight decrease in effect size as a function of increasing year ($b = -.01$). When the data from 1979 were excluded, significant heterogeneity was no longer present.

***DAT-VR.*** The summary effect size indicated a moderate male advantage on this test. It appeared that the significant heterogeneity in the DAT-VR group was due to one sample (Study 314C; see Figure 5) that demonstrated a moderate sex difference favoring females, contrary to the other 13 studies. It is unclear why this sample demonstrated a contrary result; there did not appear to be any systematic differences in sample characteristics or administration between this sample and the remainder. All studies identified using the DAT-VR were conducted on Spanish populations, with the involvement of the same researcher. Therefore, study methods and population are very similar across samples, with the possible exception of studies 312, 319, 322 and 323, which, although published by the same research team, utilized data from standardization samples and university applications. The summary effect for university samples on the DAT-VR did not demonstrate significant heterogeneity, although this may due to the smaller number of studies concerning only university students. This summary statistic was higher than the overall DAT-VR summary statistic, and may be best interpreted as the effect size in Spanish university populations specifically.

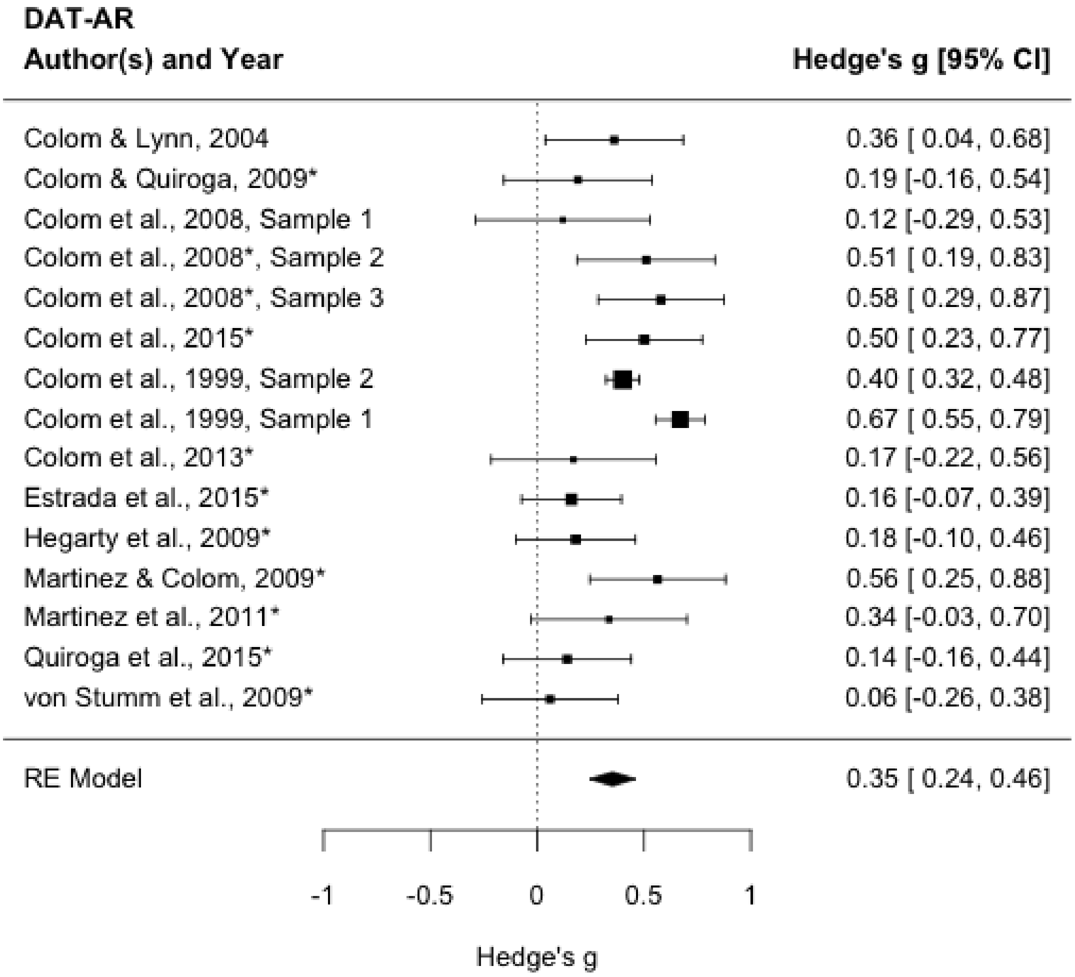


*Figure 4.* Forest plot of effect sizes for the Differential Aptitude Test – Abstract Reasoning. * indicates university student sample. Includes data from 1979 (Study 322B; Colom et al., 1999) that was excluded from analysis presented in Table 4.

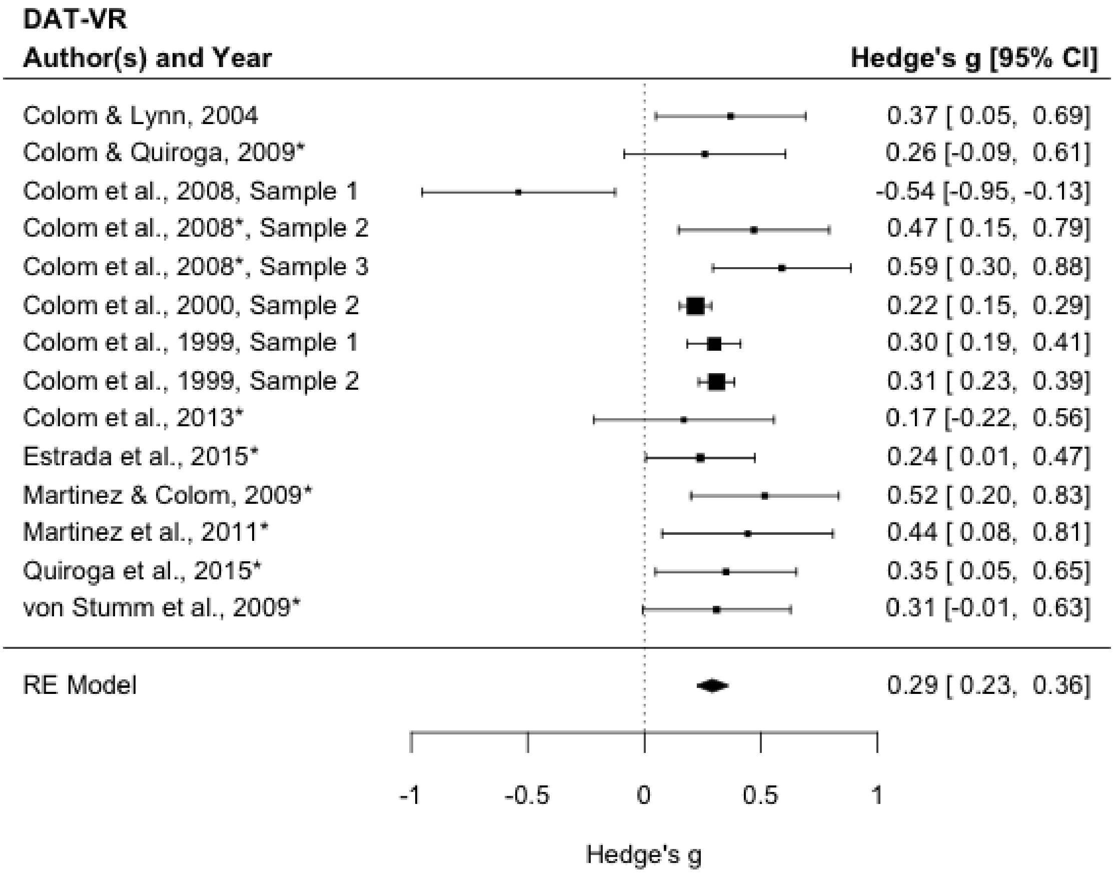


*Figure 5*. Forest plot of effect sizes for the Differential Aptitude Test – Verbal Reasoning. * indicates university student sample.

**PMA-R.** Contrary to the other measures, the PMA-R summary effect showed a female advantage. Although there was one data point (g = 0.2, Study 316A) in the PMA-R group demonstrating an effect size in the opposite direction to the remainder, exclusion of this study did not account for the significant heterogeneity in effect sizes. Visual inspection of the data shows that there was substantial variation in effect sizes reported across all studies (Figure 6). Again, all but one study concerning the PMA-R were conducted on Spanish populations, with the involvement of the same researcher and near-identical methods and procedures, with the exception of the standardization data and university applicants. It therefore it appears that there was little consistency in the strength of the male-female difference, although a female advantage was typically reported.

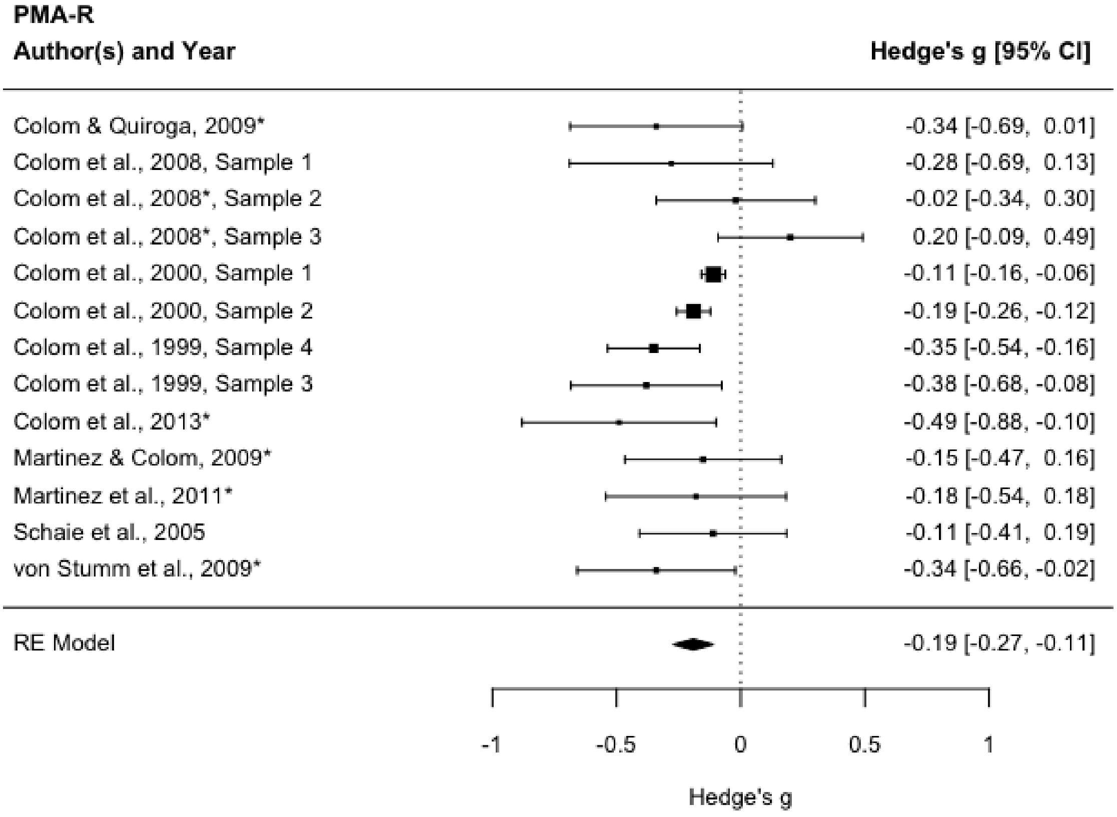


*Figure 6.* Forest plot of effect sizes for the Primary Mental Abilities - Reasoning. * indicates university student sample.

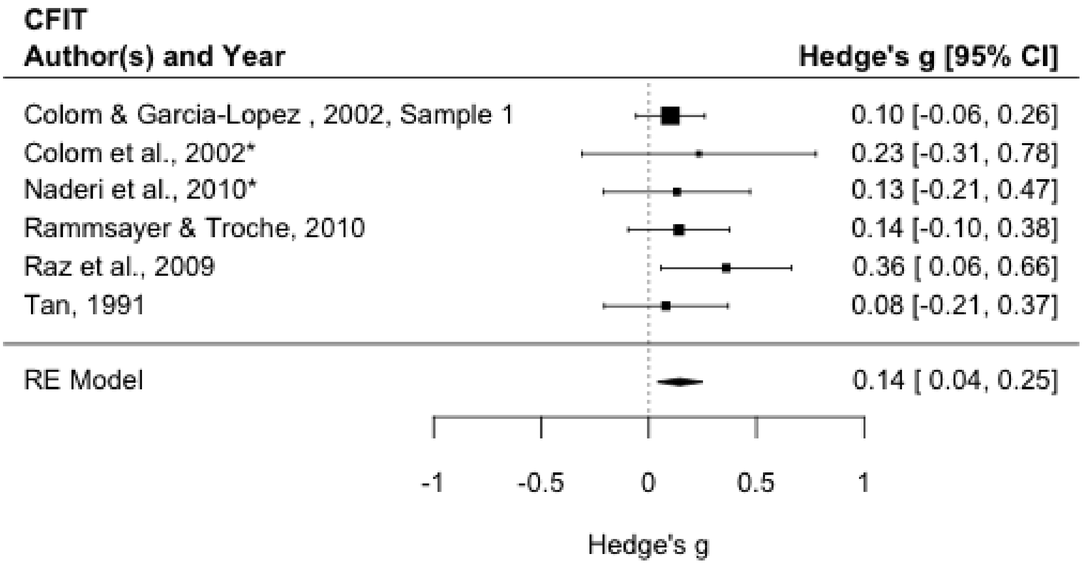


*Figure 7.* Forest plot of effect sizes for the CFIT. * indicates university student sample.

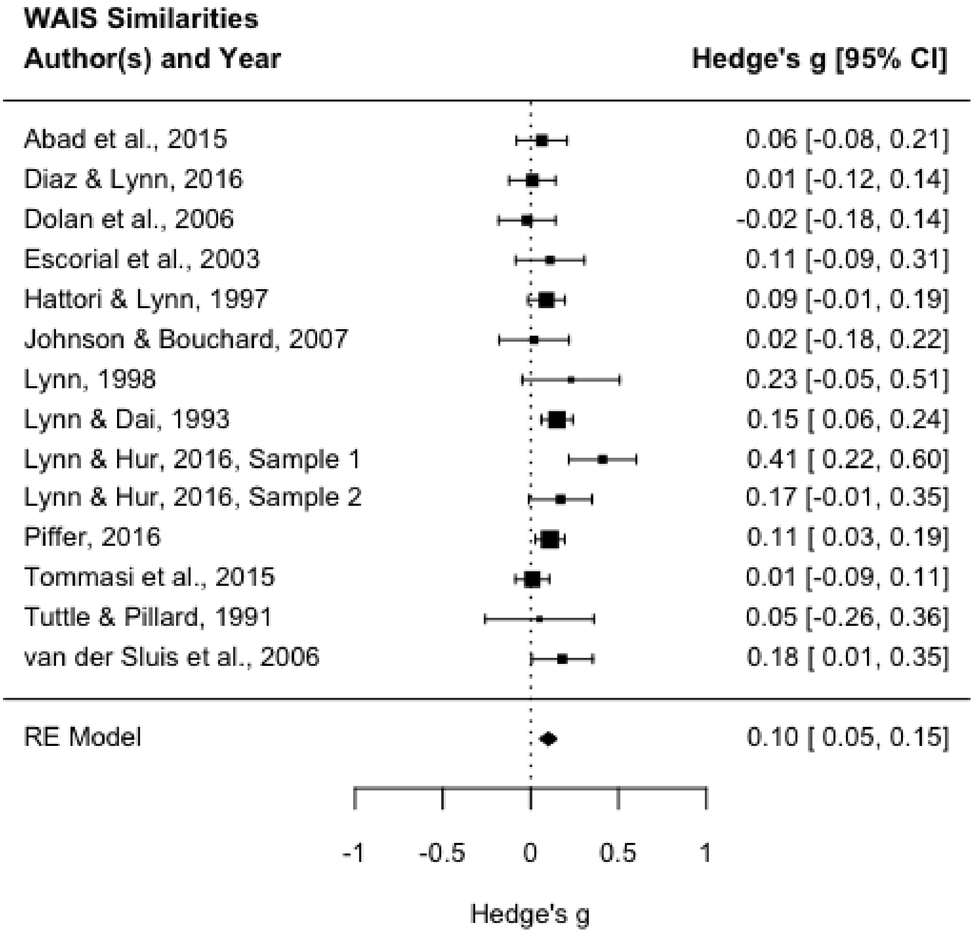


*Figure 8.* Forest plot of effect sizes for the WAIS - Similarities.

## Sensitivity Analysis and Publication Bias

Several studies identified reported that there was or was not a significant difference between males and females in the targeted measure, but did not report statistics required for calculating an exact effect size. These studies were excluded from analysis, as per inclusion criteria. However, sensitivity analysis was run including these studies, coded with an effect size of $g = 0$ to determine the effect this may have had on results. Results were largely the same, with all summary effect sizes within ± .02 (except for Alphabetic Series and PMA-R where the effect size decreased by .05 and .03, respectively). Because the effect size estimate of $g = 0$ for these studies was likely an underestimate, and results did not differ substantially by excluding these studies, results were reported using only those studies enabling calculation of an exact effect in order to provide a more precise estimate.

Sensitivity analysis was also conducted to determine the effect of including those studies with severely unequal proportions of males and females, based on moderator analysis indicating that this may affect the summary effect size. When analysis was run including only those studies with reported data from samples consisting of at least 25% males or females, the results were again largely similar with summary effect sizes within ± .02. The exception was Verbal Analogies, the DAT-AR and DAT-VR summary effects for university samples and the overall DAT-VR summary effect, which decreased by .07, .15, .09 and .07, respectively. However this was likely due to a substantial decrease in the number of data points, and consequent unreliability of the summary estimates; only three studies concerning the DAT-AR and two concerning the DAT-VR remained in the university samples. Given that the inclusion of these studies did not affect the majority of summary effects reported here, these studies were included in the overall analysis.

To examine the effect of random selection of studies concerning the same sample on the moderation effect of stimulus, 10 different random selections were tested. Of those samples that contained at least 3 cases of each stimuli-type category, the average effect size for verbal was .43 lower than figural, with an average *p*-value of .013 (range < .001 - .078). In all cases, the difference between figural and alphabetic was at least 0.39, *p* < .001 (average effect size of 0.56). This suggests the difference between figural and verbal stimuli, when accounting for item type, is not as robust as that between figural and alphabetic, although verbal is still lower.

Finally, publication bias in the identified studies was assessed**.** Figure 9 displays the associated funnel plot. A regression test (Egger, Smith, Schneider & Minder, 1997) indicated that there was no significant asymmetry present (*z* = -0.22, *p* = .828).

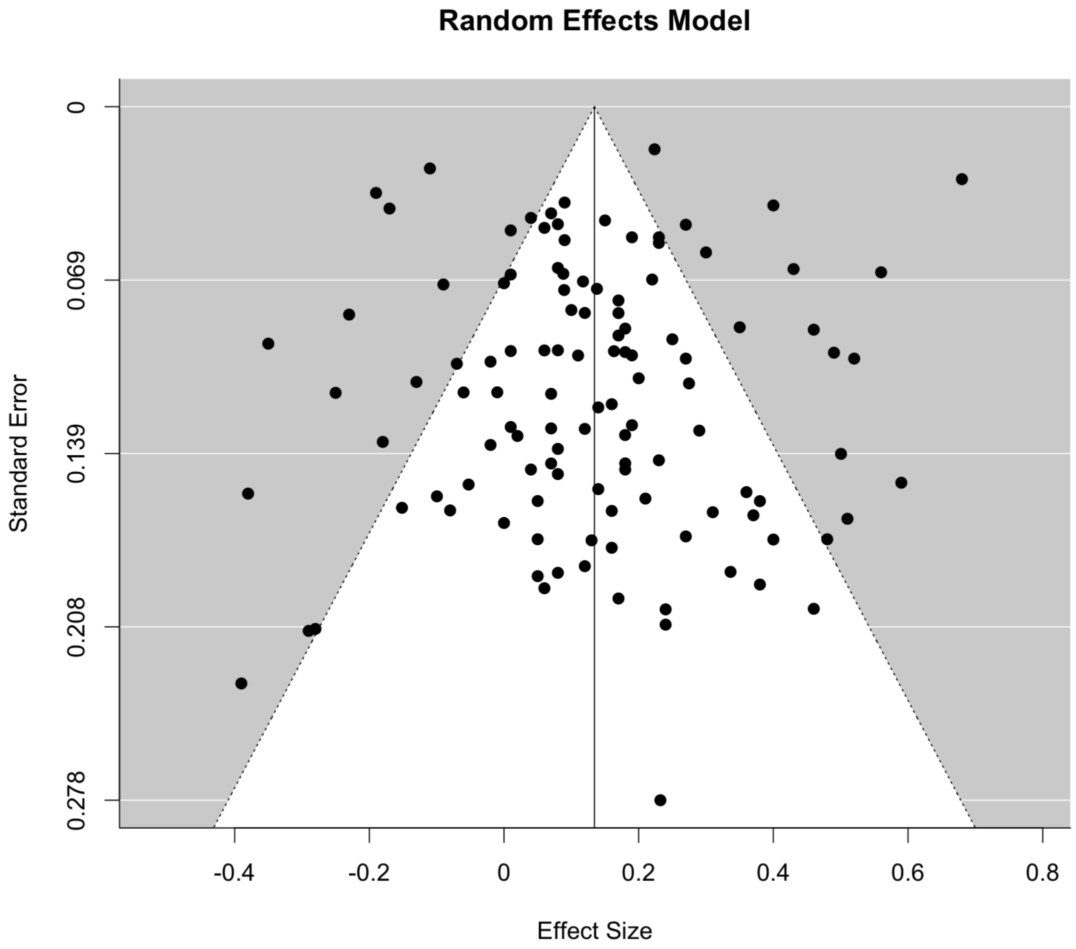


*Figure 9.* Funnel plot.

## Discussion

The results of this research synthesis suggest that there may be no consistent difference between males and females in scores on inductive reasoning measures in general. Although results indicated a male advantage on the majority of tests considered here, substantial variability in the magnitude, and even direction, of the effect size was found both between and within test type, stimuli and item type used, and specific test. The overall summary effect for inductive reasoning tests (including verbal classification and verbal analogies) was *g* = 0.13, 95% CI (.10, .17). However, a breakdown of the different tests indicates that this is not an accurate representation of the sex differences; many tests showed a weak-to-moderate male advantage (RPM tests, figural tests), some showed a slightly stronger male advantage (Figural series, DAT-VR), others showed a small difference (Verbal Classification), and yet others showed a moderately weak female advantage (PMA-R).

### Item Stimuli and Item Type

Although the type of stimulus was found to significantly moderate the sex difference in test scores, it was not able to explain all heterogeneity in scores. The effect of figural stimuli is noteworthy, however, because many commonly used tests of inductive reasoning are figural. If tests using figural stimuli are likely to favor males, this may be problematic, particularly if this is because of test characteristics rather than the underlying ability that we seek to measure, inductive reasoning. Research does indeed suggest that there is no sex difference on latent inductive reasoning ability (Arendasy & Sommer, 2012), or on the higher-order latent Gf ability (Keith et al., 2008; Lakin & Gambrell, 2014), so it is plausible that test characteristics are causing any apparent sex differences.

One suggested explanation for the male advantage on figural tests is the visual complexity of the items, which may require a certain level of Gv. The only comparison among tests using the same type of question but different stimuli that can be made in the current study is between figural series tests (DAT-AR) and alphabetic series tests (PMA-R). There was a large difference in the effect sizes found for these two tests, but a comparison of further item and stimuli types would be needed to make any stronger conclusions.

**Individual Tests**

Results for figural tests in general support Lynn and Irwing (2004b) and Irwing and Lynn's (2005) contention of a male advantage. However, the effect size found for the RPM tests specifically was slightly lower than the Cohen's *d* of 0.22-0.33 reported by Irwing and Lynn and Lynn and Irwing, ranging from Hedges' *g* of 0.10-0.26 in the current study. Figural tests utilizing various different item types (i.e. not solely matrices) showed a similar effect magnitude to the RPM tests at *g* = 0.21. The DAT-AR demonstrated the strongest male advantage of all figural tests (*g* = 0.31-0.32). Given the lack of homogeneity found in many summary effect sizes however, caution should be exercised when comparing the relative strength of effects.

With regard to the DAT-AR, Feingold (1988) reports no sex difference in four different US DAT standardization samples ranging from the years 1947 to 1980. This is a noteworthy difference to the current results. It is possible that the differences are due to age effects (for example, Lynn and Irwing [2004b] reported an increasing male advantage on the RPM with age), or the relative selectivity of samples considered. Feingold's sample concerned school students from grades 8 to 12. In the present study, standardization data from Spain was included in the DAT analyses, however the samples were slightly older than Feingold’s sample, with a mean age of 18. It was also not surprising to identify data collection date as a significant moderator of the DAT-AR effect size; Colom, Quiroga and Juan-Espinosa (1999) provided evidence for this in their study. However, it was interesting that no other test data showed a significant moderation effect of data collection date, particularly the DAT-VR analysis which also contained data collected in 1979. It may be that the effect of data collection date on the DAT-AR is particularly large due to the specific Spanish context.

Although most tests demonstrated a small male advantage, the PMA-R demonstrated a female advantage and the WAIS Similarities demonstrated a negligible difference. The PMA-R results are in stark contrast to most of the other measures. It appears that, although this test utilizes a “series” item type, and is typically accepted as a measure of Gf, there may be something quite unique about it. This finding deserves further investigation, and should be kept in mind when using this measure and interpreting the associated results. Additionally, this finding may extend to other alphabetic series tests; however, there were not sufficient data points for other alphabetic series tests to analyze them separately. It may be that, similar to what is argued regarding the role of Gv in figural tests, some other ability is involved in tests using alphabetic stimuli. A likely candidate for this would be Gc. However, verbal analogies and verbal classification did not show a similar result, even though it is commonly argued that they also involve some element of Gc. Furthermore, it is unclear how this notion fits in with the idea that females show an advantage on specific Gc abilities such as verbal ability and word knowledge, but males show an advantage on verbal analogies (Hyde & Linn, 1988). If both verbal analogies and alphabetic series tests involve reasoning, the argument that the involvement of reasoning ability in verbal analogies is responsible for the male advantage (e.g. Steinmayr, Beauducel & Spinath, 2010) does not hold.

## Limitations

One limitation of the current study was the use of unrepresentative samples in many of the studies identified. Males and females may be unequally recruited into university samples generally (Hunt & Madhyastha, 2008), and many of the samples here were university samples. With regard to the overall summary effect sizes reported, it should be kept in mind that few of these studies report results from truly representative samples. However, there did not appear to be obvious or systematic differences between the results of those studies with unrepresentative samples and those with more representative samples. Of the individual tests considered, most were largely university samples (with the exception of the SPM, WAIS Similarities and WAIS Matrix Reasoning). Therefore, these results can largely be interpreted in reference to university student samples. It should also be noted that one of the main aims of the current study was to investigate the role of stimuli and test item type in the magnitude and existence of sex differences on these measures; results indicated that there were differences in the effects of sex even within university samples. These results indicate that we should be cautious in using only one measure of inductive reasoning, specifically when considering sex differences in an ability rather than a specific measure, but also more broadly in measuring inductive reasoning ability, until the cause of these variations is understood.

A further limitation of this study could be the use of a random searching strategy rather than an exhaustive one. Reasons for this strategy have been outlined, and there is no evidence for any bias in the different samples. However, it would be desirable to expand on the current study with an exhaustive look at all published literature concerning the included measures to confirm the present results.

Finally, we must note that a large majority (if not all) of the studies concerning the DAT-VR, DAT-AR and PMA-R come from a research team including the same author

(Roberto Colom). A moderator analysis (involvement of Colom coded as a yes/no categorical variable) suggested that this was not a moderator of the effect size for the overall dataset, with research associated with Colom only very slightly favoring females ($Q(1)$ 0.04, $p = .845$, $b = -.009$). Additionally, this variable was not a significant moderator of APM effect size ($Q(1) = 1.37$, $p = .242$, $b = 0.07$) which was the only singular measure for which there were enough studies identified with and without the involvement of this author to allow comparison.

**Implications for Future Research and Practice**

The results of this study suggest that males obtain higher scores on figural measures of inductive reasoning, and furthermore that the use of a figural stimulus accounts for some of this difference. Although males also obtained higher scores on some verbal measures, particularly the DAT-VR, the sex difference was not as consistently large, particularly after accounting for item type. Given the lack of evidence of a male advantage at the latent ability level (Arendasy & Sommer, 2012; Keith et al., 2008; Lakin & Gambrell, 2014), this is problematic for the measure of inductive reasoning using manifest scores. This is also problematic because the large majority of measures of inductive reasoning utilize figural stimuli, as evidenced by the vast majority of measures included in the present study utilizing this particular stimulus type. Furthermore, although sex differences are not indicative in and of themselves of the ability measured, the fact that the PMA-R demonstrated such a different result to the figural inductive reasoning measures indicates that it may measure something slightly different. The variation in the size of the sex differences across the different measures indicate that caution should be used in interpreting individual scores, or differences between individuals based on the result of a single measure.

Future research should aim to determine the cause of the differences found between figural inductive reasoning tasks and letter series tasks; understanding this may help to further understanding of what these tests measure. Future research into alternative test formats for inductive reasoning measures should also be conducted. Additionally, it should be noted that alternative models of intelligence, such as the Berlin model of Intelligence Structure (BIS; Jäger, 1982; Süβ & Beauducel, 2005), do make provisions for the inclusion of different stimuli to measure the same abilities; however, it is currently unclear what may be equivalent to inductive reasoning. Based on the distinction between inductive, deductive and quantitative reasoning, figural reasoning appears to be the strongest contender, but what then of letter series tasks?

# Chapter 6: Paper 4

## Preamble

This study was not initially planned as part of the current thesis. However, during the search for standardisation data to be used in Paper 3, the opportunity to have a closer look at a large sample of Australian data for the GRT2 came about. Although this study is not directly related to the aims of the thesis, there is sufficient overlap in the content considered here for it to be included. Specifically, the interest in analysing these data came from a desire to investigate to what extent the so-called content facets (abstract, verbal and numeric) influence the concept measured. The test purports to measure "general reasoning" and as such, the differentiation of the abstract, verbal and numeric reasoning subtests was of interest. Additionally, we were interested in determining whether a sex difference, similar to that found on the RPM tests, existed in the GRT2.

Because little work had been conducted on the psychometric properties of the GRT2 within the Australian population, the first step was to examine the factor structure of the test. Following this, measurement invariance, differential item functioning and the role of stimulus and item type could be assessed.

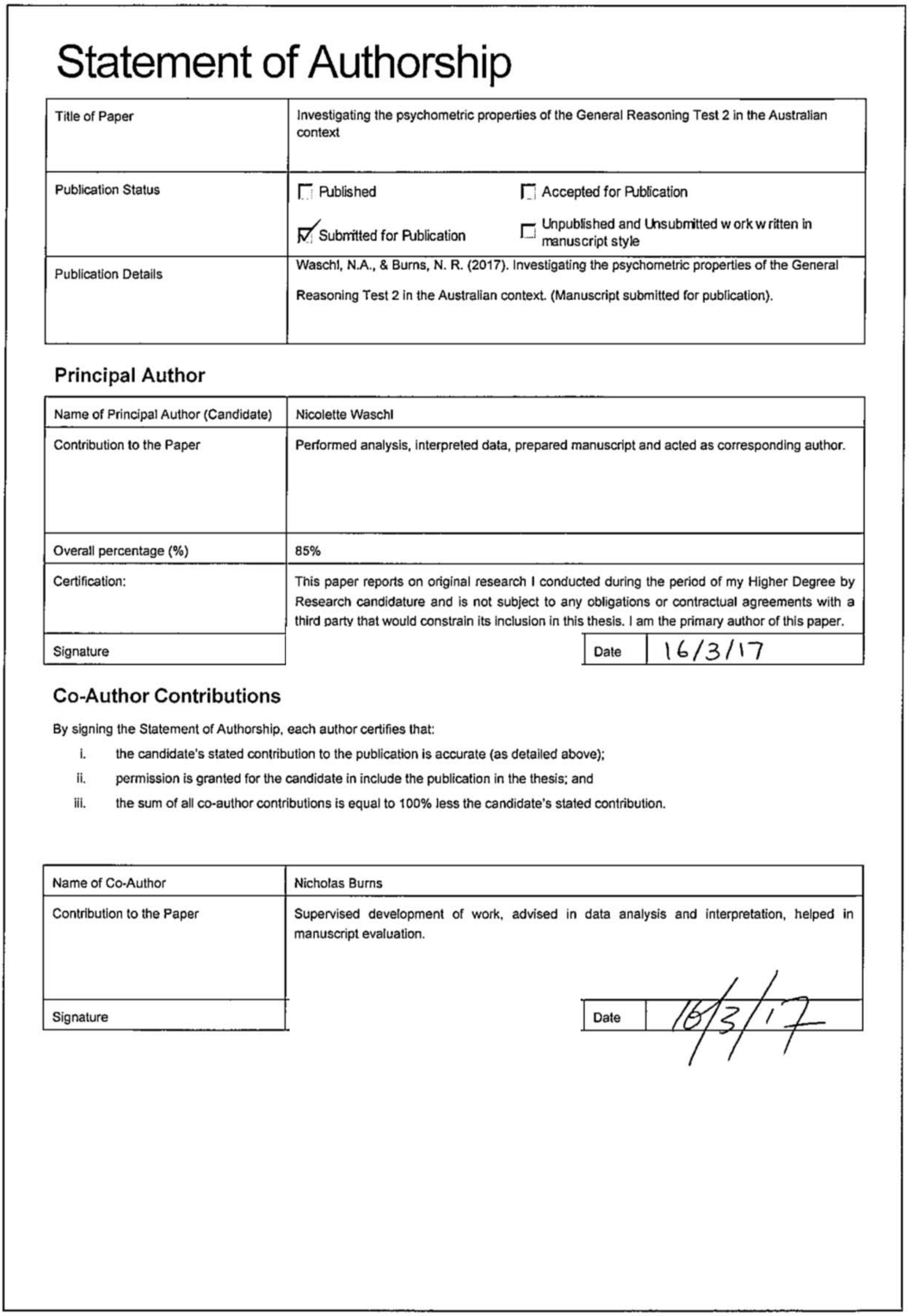
In the case of Paper 4, classical test theory methods were used because item response theory is largely a confirmatory analysis technique, relying on previous identification of a single or multiple unidimensional scale(s), and exploration of the factor structure of the GRT2 required the use of exploratory techniques in the first instance. The sample was divided into two random groups. Analysis of the first group used EFA and ESEM to examine the factor structure and sex differences, while the second group was used to confirm the factor structure and sex differences using CFA and MGCFA. MTMM modeling was used to assess validity, with a particular focus on examining the role of the content facets.

# Paper 4

# Investigating the psychometric properties of the General Reasoning Test 2 in the Australian context

Waschl, N.A., & Burns, N. R. (2016). Investigating the psychometric properties of the General Reasoning Test 2 in the Australian context. (Manuscript submitted for publication).

***Note.*** Appendix materials can be found in Appendix D.



# Statement of Authorship

| | |
|---|---|
| Title of Paper | Investigating the psychometric properties of the General Reasoning Test 2 in the Australian context |
| Publication Status | □ Published □ Accepted for Publication<br>☑ Submitted for Publication □ Unpublished and Unsubmitted w ork w ritten in manuscript style |
| Publication Details | Waschl, N.A., & Burns, N. R. (2017). Investigating the psychometric properties of the General Reasoning Test 2 in the Australian context. (Manuscript submitted for publication). |

## Principal Author

| | |
|---|---|
| Name of Principal Author (Candidate) | Nicolette Waschl |
| Contribution to the Paper | Performed analysis, interpreted data, prepared manuscript and acted as corresponding author. |
| Overall percentage (%) | 85% |
| Certification: | This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper. |
| Signature | Date 16/3/17 |

## Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

| | |
|---|---|
| Name of Co-Author | Nicholas Burns |
| Contribution to the Paper | Supervised development of work, advised in data analysis and interpretation, helped in manuscript evaluation. |
| Signature | Date 16/3/17 |

## Abstract

**Objective:** The General Reasoning Test 2 (GRT2) has been used extensively for occupational testing in Australia over the last decade. Existing standardization data from the UK provide evidence for the validity and reliability of this test. However, whilst Australian norms exist, a thorough psychometric investigation of this test in the Australian population has yet to be conducted. Therefore, the objective of this study was to evaluate the psychometric properties of this test in an Australian sample.
**Methods:** This paper used data obtained from occupational testing over the last decade to investigate the psychometric properties of the GRT2. Exploratory and confirmatory factor analyses were used to evaluate the factor structure of the test. Differential item functioning and latent mean differences were examined across sex and age. The presence of method factors was also investigated.
**Results:** Results suggest that a two-factor (Gf-Gc) solution best represents the structure of the GRT2. Several verbal reasoning items were found not to load on either factor. The test was found to be largely measurement invariant across sex; however, sex differences favouring males were identified at the latent factor level.
**Conclusions:** The proposed Gf-Gc structure of the GRT2 was confirmed. However, some verbal reasoning items do not appear to function well in the Australian population. Test users should be aware of some sex and age differences in items and latent means.

*Keywords*: Differential item functioning; Factor structure; General Reasoning Test 2; Psychometric properties; Sex differences

For many years, cognitive ability tests have been a popular method of assessing job applicants. Although many other methods can be used for this purpose, cognitive ability testing remains one of the most valid, predicting around a quarter (16-36%) of the variance in job performance and performance in job training programs (Hunter, 1986; Schmidt & Hunter, 1998, 2004).

The General Reasoning Test 2 (GRT2) is a measure of general reasoning ability developed to assess job applicants. This test was designed for use with the general population, for assessment related to job roles requiring an average level of reasoning ability (Psytech International, n.d.), and is based on Cattell's (1971, 1987) fluid and crystallized intelligence. Fluid intelligence (Gf), as defined by Cattell, is a biological or innate capacity for solving novel, previously unlearnt problems, while crystallized intelligence (Gc) involves the ability to use knowledge and experience. Both are core cognitive abilities.

The GRT2 consists of three subtests: verbal reasoning (VR), numerical reasoning (NR) and abstract reasoning (AR). Each subtest is presented in a multiple-choice format and is timed (VR 8 min; NR and AR 10 min each). Each subtest consists of several different item types, to ensure that the construct being measured is not too narrow (for example, the verbal subtest consists of analogies, vocabulary, and classification items). The GRT2 is based on the Alice Heim series of tests (Heim, 1970), which also consist of three subtests (verbal, numerical and perceptual). Individualised reports are provided which report stanine scores for each subtest, for general mental ability (*g*), and for Gc and Gf. Information on how the *g*, Gc and Gf scores are calculated was not available to us.

The GRT2 was developed in the UK, and the test manual largely presents standardisation data regarding British adults. According to these data, the GRT2 shows good reliability and validity in various occupational groups and for various outcomes (Psytech International, n.d.). Australian norms are available, and some additional research has been conducted on the psychometric properties of this test (for example, within the New Zealand population [McInnes, 2011]), but an extensive investigation of the psychometric properties within the Australian population has yet to be conducted. Since 1993, this test has been administered to hundreds of thousands of job applicants in Australia for the purpose of occupational testing. The associated data present a good opportunity to investigate the psychometric characteristics of this measure in the Australian population specifically, and also demonstrates a need for more understanding of the properties of the test within this population.

To our knowledge, little research has been conducted on the factor structure of the GRT2. This is likely because of its widespread use in applied, organisational settings, rather than academic spheres. Although the test was conceptualised to measure Gf and Gc, there is little information available regarding factor analytic evidence of this structure. Some research reports a single-factor (*g*-factor) structure in the British (Furnham, Moutafi & Paltiel, 2005; Moutafi, Furnham & Paltiel, 2005) and South African (Odendaal, 2015) populations, but other evidence points toward a two-factor (Gf-Gc) structure (McInnes, 2011). The structure of this test has not yet been explored within the Australian context and, given conflicting empirical and theoretical indications of the appropriate factor structure, this requires investigation.

Potential sex and age differences in the latent structure and on individual items are also an important consideration. Investigation of issues of measurement invariance is required to ensure equality and comparability across groups. Establishing measurement invariance ensures that the construct measured by the test is the same across different groups and, without establishing this, we cannot be confident in making comparisons across these groups. While the GRT2 may not typically be used to compare across sex or age groups directly, if the test disadvantages a particular group this would certainly be necessary to know.

Research within the broader cognitive ability literature indicates that males tend to perform better than females on tests of numerical reasoning (Keith, Reynolds, Patel & Ridley, 2008; Lakin & Gambrell, 2014). There are conflicting findings regarding tests of verbal and abstract reasoning. Females have been found to perform better on verbal tests generally (Hyde & Linn, 1988), and on verbal Gf tests specifically (Lakin & Gambrell, 2014), but there is evidence that males perform better on verbal analogies tests (Hyde & Linn, 1988). It may be that sex differences on verbal reasoning tests are dependent on the specific type of test question and item. For example, analogies items may tend to favour males more strongly than other types of verbal items, such as vocabulary and anagrams (Hyde & Linn, 1988). Given that the GRT2 combines many types of verbal reasoning items, it is unclear what to expect with regard to sex differences in this subtest. With regard to abstract reasoning, studies of sex differences in these tests have reported all possible results across several different test batteries (Colom & Garcia-Lopez, 2002; Colom & Lynn, 2004; Keith et al., 2011; Lakin & Gambrell, 2014; Lohman & Lakin, 2009). It is thus unclear whether any true sex difference exists in abstract reasoning or should be expected with the GRT2 abstract reasoning subtest.

Additionally, there is evidence for age differences in the abilities purported to be measured by the GRT2. Both cross-sectional and quasi-longitudinal studies have shown that Gc abilities, such as vocabulary, tend to increase with age until declining in late life, while Gf abilities may slowly decline from early-to-middle adulthood (Salthouse, 2016). However, some research has indicated that these differences may be quite modest (e.g., ± .02 standard deviations; Salthouse, 2012).

Australian norms indicate that males show a raw score advantage on the numerical and abstract reasoning subtests of the GRT2, and that age differences exist in raw scores on the abstract reasoning subtest (Genesys Australia, 2012). Differential item functioning across sex has been examined within the British population, with results demonstrating very little evidence of item bias (Psytech International, n.d.). Nonetheless, there may be cultural influences in any sex item bias and as such it is pertinent to investigate this within the Australian context.

Finally, it is interesting to consider whether the type of question influences factor loadings or sex differences. As previously described, the GRT2 uses three different stimuli to measure reasoning: abstract, numeric and verbal, but also uses various item types to assess these different reasoning abilities (analogies, series, classification, vocabulary, mathematical ability). It is unclear how the type of question might influence the factor loadings of certain items; however, it is likely that items measuring vocabulary and mathematical ability, for example, might be found to load on a Gc factor, while analogies and series items (whether they be abstract, numeric or verbal) may be more likely to load on a Gf factor, being the types of items that typify Gf (Carroll, 1993). Because the GRT2 uses such a variety of item types, it presents a unique opportunity to address this question.

This consideration poses two elements of investigation. The first entails an assessment of the construct validity of the GRT2 to determine convergent and discriminant validity of method (item or stimulus type) and trait (ability) factors, and to explore the existence of method effects. The second is more tangential to the GRT2 measure itself and involves exploring the relationship between the stimulus and the item type. Many tests of reasoning use only one item type and one stimulus type (e.g. the Raven's Progressive Matrices) and it is therefore pertinent to consider the relative contribution of these two item elements to explaining item variance and, consequently, the utility of including multiple item and stimulus types. Both of these questions of construct validity and the role of stimulus and item type can be assessed using multitrait-multimethod models, with different emphasis on item or stimulus type versus ability and item type versus stimulus type.

The first aim in examining the psychometric properties of the GRT2 was to ascertain the factor structure of this test. Since the test was designed to measure Gf and Gc, but some evidence indicates a single-factor structure, exploratory factor analysis was initially used, followed by confirmatory factor analysis. The second aim of this study was to investigate the effects of sex and age at both the latent ability level and the item level. A third aim was to assess whether the type of question influenced factor loadings or any sex differences identified in items, and associated with this, construct and discriminant validity related to the presence of method effects.

## Method

### Participants and Data Preparation

Participants were drawn from a sample of 21,944 Australian adults who had completed the test as part of occupational testing between the years 2001 to 2015. There were substantial missing data regarding the demographic characteristics (age, education) of this sample. Of the cases available, only data for those who answered every test item was included in analysis. This was due to the method of recording responses, which did not distinguish between an unanswered question and a question answered incorrectly, but did give the total number of answered questions. GRT2 items do not increase in difficulty and test-takers are able to skip questions. As such, there was no way of determining which questions were not answered and which were answered incorrectly unless all items were completed. Therefore the cases used in the present analysis numbered 3494.

The dataset was randomly divided into two samples, each consisting of 1747 individuals. Group 1 was used to conduct the exploratory analysis, while Group 2 was used to conduct the confirmatory analysis. Group 1 consisted of 1201 males and 539 females (7 did not report gender). The mean age was 30.3 years (*SD* = 11.4). The mean total score of this group was 67.2 (*SD* = 12.2; potential maximum score was 85). Group 2 consisted of 1231 males and 512 females (4 did not report gender). The mean age was 30.6 years (*SD* = 11.1). The mean total score of this group was 67.1 (*SD* = 12.2). The total score did not statistically differ between groups, $t(3492) = 0.29$, $p > .05$, and neither did age $t(2741) = 0.70$, $p > .05$. The proportion of males and females in each group was also not statistically significantly different, $\chi^2(1) = 1.06$, $p > .05$.

The mean score of both Group 1 and Group 2 for each subtest was higher than both the 2007 and 2012 Australian norms (see Table 1). This is likely due to the fact that respondents who answered all questions are likely to have scored higher than the norm sample. Supporting this, although the scores of Groups 1 and 2 on each subtest were not statistically significantly different from each other, they were statistically significantly higher than subtest scores of individuals from the initial sample not included in Groups 1 and 2 (AR: $F(2, 21941) = 638.92$; NR: $F(2, 21941) = 1633.24$; VR: $F(2, 21941) = 703.73$; all $p < .001$; see Table 1 for means and standard deviations).

Table 1

*Subtest means for Group 1 and 2 compared to 2007 and 2012 Australian norms*

| Subtest | Group | Possible Max | Mean (SD) | 2007 Mean | Mean Diff | *d* | 2012 Mean | Mean Diff | *d* |
|---|---|---|---|---|---|---|---|---|---|
| AR | Group 1 | 25 | 20.09 (4.1) | 16.7 (4.9) | 3.4 | 0.75 | 17.4 (4.7) | 2.7 | 0.61 |
| | Group 2 | | 20.10 (4.2) | | 3.4 | 0.75 | | 2.7 | 0.61 |
| | Remainder of sample | | 17.05 (4.7) | | 0.4 | | | 0.4 | |
| NR | Group 1 | 25 | 20.54 (5.1) | 14.9 (5.6) | 5.6 | 1.05 | 15.7 (5.7) | 4.8 | 0.89 |
| | Group 2 | | 20.51 (5.1) | | 5.6 | 1.05 | | 4.8 | 0.89 |
| | Remainder of sample | | 14.88 (5.4) | | 0 | | | 0.8 | |
| VR | Group 1 | 35 | 26.57 (4.7) | 23.0 (5.5) | 3.6 | 0.70 | 23.6 (5.1) | 3.0 | 0.61 |
| | Group 2 | | 26.48 (4.5) | | 3.5 | 0.69 | | 2.9 | 0.60 |
| | Remainder of sample | | 22.87 (5.4) | | 0.1 | | | 0.7 | |
| *N* | | | | 11673 | | | 9150 | | |

*Note*. Group 1 and 2 *N* = 1747; Remainder of (initial) sample *N* = 18450. AR = Abstract Reasoning; NR = Numerical Reasoning; VR = Verbal Reasoning.

## Results

### Exploratory Factor Analysis

**Factor extraction criteria.** Parallel analysis of both principal components and factors and inspection of the corresponding scree plots using the psych package (Revelle, 2015) in R (R Core Team, 2015) was used to determine the number of factors to extract. The results from parallel analysis and scree plots of the factors suggested the presence of over-factoring; examination of the corresponding solutions confirmed this. Therefore, parallel analysis and scree plots of principal components were used.

Factor extraction criteria suggested extracting one factor for each of the abstract and numerical reasoning subtests, three for the verbal reasoning subtest, and four for the test as a whole. Although the first principal component accounted for a substantially greater proportion of variance than the subsequent components, exploration of a one-factor solution indicated substantial unexplained variance (approximately 70%). Therefore, this structure was not further tested. According to Kline (1994), the finding of a single large general factor in principal components analysis can be an artifact of the algebraic procedures and should not be given undue attention before rotations are performed. Factor structures are more accurately interpretable after factor rotations, when the variance from the general factor is redistributed.

**Exploratory factor analysis (EFA).** EFA was conducted in Mplus 7 (Muthen & Muthen, 1998-2012) using the Weighted Least Squares Mean and Variance adjusted (WLSMV) estimator with geomin and bi-geomin rotation. Analyses were performed separately for each subtest, and for the whole test. When analysis was conducted separately on each subtest, the one-factor solutions for abstract and numerical reasoning demonstrated good fit (see Appendix I). Items 3 and 6 of the abstract reasoning subtest had low factor loadings, but all other items showed acceptable loadings on their respective factors.

The factor structure of the verbal reasoning subtest was more complex, with five items found not to load substantially on any of the three factors (i.e., highest loading < .25). These items also had low item-total correlations, ranging from *r* = .06 to .35, and therefore were iteratively removed from analysis. These items were VR2, VR3, VR8, VR19 and VR28. Factor extraction criteria subsequently indicated two factors. All items retained showed acceptable loadings on one of the two factors (Appendix I).

While the parallel analysis suggested the extraction of four or six factors for the test as a whole, there were theoretical reasons to believe that a two-factor (Gf-Gc) solution may adequately explain the variance in this test. Given this, four-, six- and two-factor solutions were all inspected. The six-factor solution showed clear over-extraction. The four-factor solution showed an adequate fit, and produced an abstract reasoning factor, a numerical reasoning factor and two verbal reasoning factors. However, when the two-factor solution was fit to the data, the items comprising the two verbal reasoning factors loaded onto the abstract and numerical reasoning factors, respectively, in a similar way to when the verbal reasoning items were considered separately.

One issue with parallel analysis of dichotomous data is the over-extraction of factors, particularly when proportions are unequal and factor loadings low (i.e. around .50; Weng & Cheng, 2005). Therefore, in the interests of parsimony, and due to the fact that the two-factor solution painted a similar picture to the four-factor solution, as well as for theoretical reasons, the two factor solution was preferred.

Table 2 presents the factor loadings for this solution. The numerical reasoning items consistently load on a single factor (Factor II), with the exception of items 5 and 6, which both show a cross-loading with Factor I. Similarly, the abstract reasoning items fairly consistently loaded on Factor I, with some additional cross-loadings on Factor II. Abstract reasoning item 4 did not show a significant loading on Factor I. The verbal reasoning items loaded onto both factors. Some verbal reasoning items demonstrated substantial cross-loadings; however the majority did show a salient loading on either Factor I or Factor II.

Factor I was interpreted as Gf, while Factor II was interpreted as Gc. The loadings of the abstract reasoning items on Factor I and numerical reasoning items on Factor II support this interpretation; the division of loadings from the verbal reasoning subtest is more complex to interpret. Table 3 presents the verbal reasoning items loading onto the Gc and Gf factors. There are four different types of verbal reasoning item in the GRT2: classification, antonym, synonym, and analogy. Classification and analogy items typify reasoning items (Carroll, 1993), while antonym and synonym items may be more strongly related to vocabulary, a Gc ability. Table 3 shows that classification and analogy items tended to load onto Factor I, while synonym and antonym items tended to load onto Factor II. In fact, 11/13 (85%) verbal reasoning items on Factor I are classification or analogy items. Although the split between analogy/classification items and synonym/antonym items is approximately 50/50 on Factor II (interpreted as Gc), this is not entirely surprising. Some of these analogy items can be seen as requiring more cultural knowledge (e.g. understanding the relationship between colours and moods, job roles etc), and therefore may be more highly related to Gc.

Table 2

*Factor loadings and communalities for the two-factor solution (Gf/Gc) GRT2 test (all subtests, excluding some verbal reasoning items)*

| | Factor Loadings | | | | Factor Loadings | | | | Factor Loadings | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Item | I | II | $h^2$ | Item | I | II | $h^2$ | Item | I | II | $h^2$ |
| AR1 | **.47** | | .25 | NR5 | **.37** | **.35** | .44 | VR8 | | | |
| AR2 | **.67** | | .50 | NR6 | **.37** | .21 | .29 | VR9 | **.34** | | .14 |
| AR3 | .28 | | .07 | NR7 | | **.57** | .42 | VR10 | .24 | .21 | .17 |
| AR4 | | **.38** | .18 | NR8 | | **.52** | .43 | VR11 | **.47** | | .27 |
| AR5 | **.35** | | .19 | NR9 | | **.38** | .24 | VR12 | **.53** | | .43 |
| AR6 | .29 | | .09 | NR10 | .21 | **.41** | .33 | VR13 | | **.37** | .11 |
| AR7 | **.55** | .24 | .55 | NR11 | | **.69** | .47 | VR14 | | **.32** | .18 |
| AR8 | **.57** | | .23 | NR12 | | **.58** | .50 | VR15 | .21 | .25 | .18 |
| AR9 | **.37** | .21 | .29 | NR13 | | **.63** | .39 | VR16 | **.36** | | .23 |
| AR10 | **.98** | -.23 | .69 | NR14 | | **.62** | .42 | VR17 | | **.38** | .24 |
| AR11 | **.54** | | .36 | NR15 | .27 | **.44** | .44 | VR18 | .25 | | .12 |
| AR12 | **.59** | | .42 | NR16 | | **.72** | .62 | VR19 | | | |
| AR13 | **.48** | | .38 | NR17 | | **.71** | .52 | VR20 | **.36** | | .11 |
| AR14 | **.30** | | .11 | NR18 | | **.76** | .55 | VR21 | .28 | .26 | .26 |
| AR15 | **.63** | | .31 | NR19 | | **.46** | .38 | VR22 | **.37** | .27 | .35 |
| AR16 | **.93** | -.23 | .60 | NR20 | | **.74** | .56 | VR23 | **.43** | .24 | .39 |
| AR17 | **.51** | | .44 | NR21 | | **.81** | .65 | VR24 | | **.38** | .14 |
| AR18 | **.35** | | .26 | NR22 | | **.92** | .76 | VR25 | **.37** | | .16 |
| AR19 | **.42** | **.39** | .57 | NR23 | | **.74** | .55 | VR26 | | **.35** | .14 |
| AR20 | **.41** | .27 | .40 | NR24 | -.21 | **.94** | .63 | VR27 | .28 | | .17 |
| AR21 | **.42** | | .33 | NR25 | | **.73** | .39 | VR28 | | | |
| AR22 | **.47** | | .31 | VR1 | **.49** | | .22 | VR29 | | **.30** | .15 |
| AR23 | **.40** | | .24 | VR2 | | | | VR30 | | **.33** | .23 |
| AR24 | **.41** | | .27 | VR3 | | | | VR31 | | **.58** | .34 |
| AR25 | **.43** | .21 | .36 | VR4 | | .21 | .14 | VR32 | **.33** | | .18 |
| NR1 | | **.45** | .33 | VR5 | **.41** | | .26 | VR33 | | **.58** | .37 |
| NR2 | | **.71** | .65 | VR6 | **.42** | | .20 | VR34 | | **.44** | .17 |
| NR3 | | **.63** | .42 | VR7 | **.52** | | .27 | VR35 | | **.38** | .28 |
| NR4 | | **.57** | .50 | | | | | | | | |

Factor Correlation

| | I |
|---|---|
| II | .72 |

*Note.* Factor loadings < .20 not displayed. Factor loadings >.30 in bold

Table 3

*Verbal reasoning items by factor*

| Factor I | | | Factor II | | |
|---|---|---|---|---|---|
| Item Type | Percent | | Item Type | Percent | |
| Classification | 54 | 85 | Classification | 27 | 54 |
| Analogy | 31 | | Analogy | 27 | |
| Synonym | 0 | 15 | Synonym | 37 | 46 |
| Antonym | 15 | | Antonym | 9 | |

With regard to types of items in the abstract and numerical reasoning subtests, Table 4 presents the average loadings of each item type. Series and Analogies items show the highest loadings on Factor I, while Series and Mathematical ability items show the highest loadings on Factor II. It is interesting that numerical series show such a high loading on the factor interpreted as Gc; however, although “series” items commonly measure Gf, numerical series items may also involve quantitative ability, which is functionally more highly related to Gc than Gf (Schneider & McGrew, 2012). This may also indicate that the role of the question format (abstract or numeric) is more important to the ability measured than the type of item.

Given the clear presence of a strong primary factor, bi-factor EFA was also conducted to determine if this type of model might be a better fit to the data. This was done using the bi-geomin rotation. Both an oblique and orthogonal solution were considered, but were found to give similar results.

Both 3 and 4-factor bi-factor models were fitted. The 4-factor model demonstrated superior fit; however, there were few significant loadings on factors other than the general factor, and therefore this model was not considered further.

Overall, EFA results indicated that two correlated factors, interpreted as Gf and Gc, best represented the structure of the GRT2. The numerical reasoning items showed the most consistent and unique loadings on Factor II. Items from the abstract reasoning subtest also tended to consistently load on Factor I, although demonstrated some cross loadings with numerical reasoning. The factor structure of the verbal reasoning items was less clear, with some items showing no demonstrable loading on any factor.

Table 4

*Average loadings of different item types on Factors I and II*

| Abstract Reasoning | | Numerical Reasoning | |
|---|---|---|---|
| Type | Average Loading on Factor I | Type | Average Loading on Factor II |
| Series | .62 *(.25)* | Series | .67 *(.17)* |
| Analogy | .47 *(.10)* | Analogy[‡] | .53 *(.23)* |
| Classification[†] | .38 *(.08)* | Classification | .42 *(.03)* |
| | | Mathematical ability | .66 *(.13)* |

[†]One item (AR4) loads on Factor II rather than Factor I and is excluded. [‡]Two items (NR5 and NR6) show higher loadings on Factor I than Factor II, but are included (Excluding these items results in *M* = .65, *SD* = .16).

**Exploratory Structural Equation Modeling (ESEM)**

Preliminary investigation of the relationship between sex, age and GRT2 factor scores and items was conducted using ESEM and data from Group 1. This was done with the test as a whole only. First, the two latent factors representative of the whole test were regressed on sex, age (centered) and age-squared. Figure 1 shows the factor scores by sex. Males demonstrated slightly higher factor scores for both factors. There was some evidence of a differential effect of age on Factor I and Factor II scores with increasing age (see Figure 2), leading to the inclusion of both the age and age-squared terms. A sex-by-age interaction was also considered, but inspection of the relevant plots indicated that it was not necessary (see Appendix II). Inclusion of the age-squared term showed that there was not much change in the effect of age for Factor I. There was, however, change in the effect of age for Factor II. The model including sex, age and age squared demonstrated good fit to the data ($\chi^2_{(3235)}$ = 4081 [normed chi-square = 1.26]; CFI = .97; RMSEA = .014). Standardised parameter estimates for Factor I were: sex = 0.294, $p$ < .001; age = 0.027, $p$ = .775; age squared = 0.090, $p$ = .346. For Factor II: sex = 0.560, $p$ < .001; age: 0.046, $p$ = .608; age squared = 0.278, $p$ = .001. The effect of age on the Gc factor was larger than the effect of age on the Gf factor.

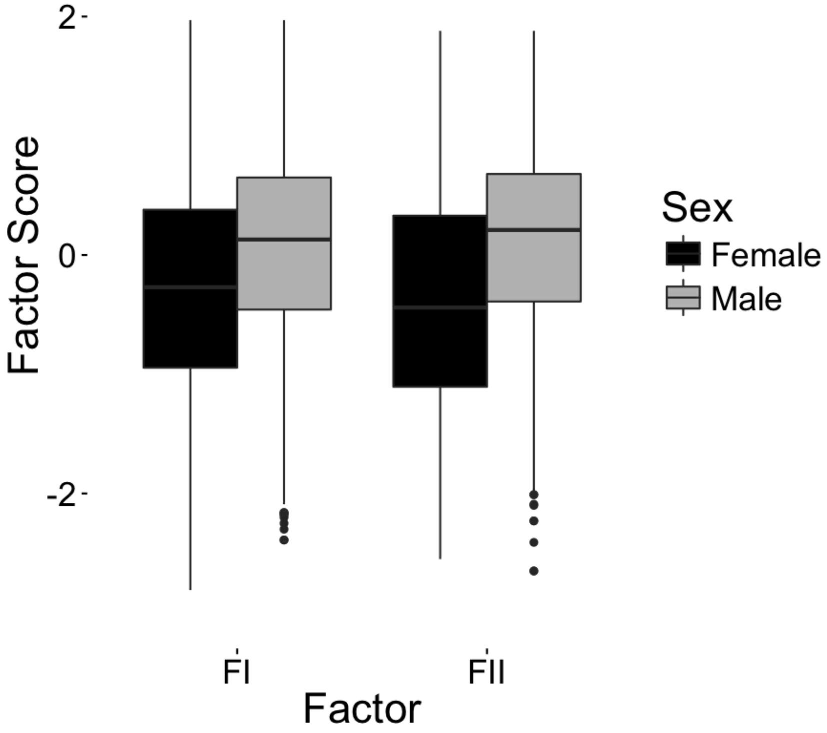


*Figure 1.* Factor scores by sex.

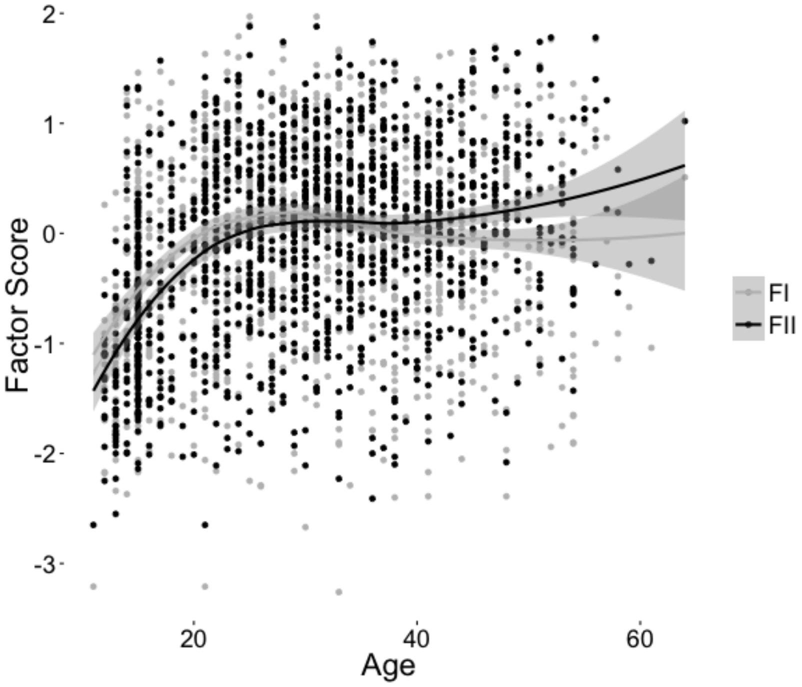


*Figure 2.* Factor scores by age. Lines indicate the smoothed conditional mean for each factor, with confidence bands indicating the 95% confidence interval.

Multiple Indicators Multiple Causes (MIMIC) modeling was used to examine differential item functioning in an ESEM framework. In order to examine Differential Item Functioning (DIF) using a MIMIC model, the covariates of sex, age and age squared were added to the model and the latent ability factors were regressed onto these covariates. Following this, paths from each of the three covariates to each item were constrained to zero, assuming no direct effect of the covariate on the item. Modification indices were examined, and paths freed in order of the highest modification indices until the chi-square critical value of 3.84 ($p < .05$) was reached. Age and age squared paths were freed in tandem. This resulted in the identification of 16 items demonstrating DIF (see Table 5). The differences were typically small with regard to sex, however they were somewhat larger with regard to age. The difference in fit between the baseline model with sex and age parameters on every item held at zero ($\chi^2_{(3235)}$ = 4081 [normed chi-square = 1.26]; CFI = .97; RMSEA = .014) and the final model ($\chi^2_{(3213)}$ = 3941 [normed chi-square = 1.23]; CFI = .97; RMSEA = .013) was not substantially different.

Table 5

*Items demonstrating Differential Item Functioning*

| Item | Parameter Estimate | |
|---|---|---|
| | Sex | Age / Age$^2$ |
| AR2 | .24 | |
| AR22 | | .10[a] / -.32 |
| NR2 | .23 | |
| NR16 | | -.33 / .51 |
| NR18 | .24 | |
| NR24 | -.23 | |
| VR7 | | -.45 / .33 |
| VR11 | -.30 | |
| VR14 | -.27 | |
| VR17 | | -.40 / .73 |
| VR26 | .24 | |
| VR29 | -.45 | |
| VR30 | -.50 | .33 / -.18[a] |
| VR31 | -.35 | |
| VR34 | -.35 | |
| VR35 | -.26 | |

*Note.* Positive values indicate differential item functioning favouring males or older individuals. [a]Parameter estimates not significant ($p < .05$).

**Confirmatory Factor Analysis**

The two-factor structure identified using EFA was tested in Group 2 using CFA. Results demonstrated a good fit to the data ($\chi^2_{(3001)}$ = 4146 [normed chi-square = 1.38];

CFI = .97; RMSEA = .02). All items showed acceptable loadings, ranging from .30 to .85, with the exception of three items (AR3, loading .26 on Factor I; VR20, loading .25 on Factor I; and VR 15 which failed to load significantly on either factor). Table 6 presents the factor loadings for the CFA analysis, and the threshold values for each item. The threshold values indicate the level of the latent factor at which an individual is expected to transition from 0 (an incorrect response) to 1 (a correct response). As can be seen from the threshold values, item AR7 appears to be particularly easy, with a threshold value of -2.05, compared to the next lowest at -1.68. It can also be seen that the difficulty of items is randomly distributed across each subtest: there is no consistent increase or decrease across the subtest.

Note that the factor correlation was much higher, at .88, using CFA procedures than using ESEM procedures, a known phenomenon when comparing the results of the two procedures. CFA uses the independent cluster model, which requires indicators to load onto only one factor, and has been said to be too restrictive in cases where indicators have secondary loadings (Marsh et al., 2009). Imposing zero factor loadings, as is the case in CFA models, can lead to distorted factors and can positively bias factor correlations (Marsh et al. 2009). It is therefore likely that the true factor correlation is lower than indicated by the CFA analysis. At a theoretical level, it is expected that verbal reasoning items would cross load on Gc and Gf factors, so this is not entirely unexpected.

However, given the high factor correlation, a one-factor CFA model was also investigated. It was found that the one-factor model demonstrated significantly worse fit than the two-factor model ($\Delta\,\chi^2_{(1)} = 160$, $p < .001$; Δ CFI = -.009; Δ RMSEA = +.002), supporting the two-factor model despite the high factor correlation.

Table 6

*Factor loadings and threshold values for GRT2 two-factor confirmatory model*

| Item | Factor I | Factor II | Threshold | Item | Factor I | Factor II | Threshold |
|---|---|---|---|---|---|---|---|
| AR1 | .51 | | -0.62 | NR19 | | .56 | -1.23 |
| AR2 | .75 | | -0.98 | NR20 | | .75 | -0.30 |
| AR3 | .26 | | -0.62 | NR21 | | .85 | -0.82 |
| AR4 | | .43 | -0.63 | NR22 | | .83 | -0.88 |
| AR5 | .44 | | -0.67 | NR23 | | .70 | -0.67 |
| AR6 | .34 | | -0.32 | NR24 | | .68 | -0.69 |
| AR7 | .70 | | -2.04 | NR25 | | .64 | -0.88 |
| AR8 | .51 | | -0.62 | VR1 | .34 | | -1.46 |
| AR9 | .59 | | -0.44 | ~~*VR2*~~ | | | |
| AR10 | .71 | | -1.26 | ~~*VR3*~~ | | | |
| AR11 | .64 | | -0.44 | VR4 | | .39 | -1.00 |
| AR12 | .57 | | -0.99 | VR5 | .59 | | -1.56 |
| AR13 | .61 | | -1.31 | VR6 | .39 | | -1.58 |
| AR14 | .38 | | -0.74 | VR7 | .49 | | 0.20 |
| AR15 | .54 | | -1.68 | ~~*VR8*~~ | | | |
| AR16 | .70 | | -1.29 | VR9 | .38 | | -1.30 |
| AR17 | .67 | | -0.76 | VR10 | | .45 | -0.67 |
| AR18 | .57 | | -0.62 | VR11 | .50 | | 0.04 |
| AR19 | .75 | | -0.90 | VR12 | .59 | | -1.15 |
| AR20 | .59 | | -1.27 | VR13 | | .30 | 0.17 |
| AR21 | .58 | | -1.19 | VR14 | | .42 | -1.10 |
| AR22 | .48 | | -1.01 | ~~*VR15*~~ | | | |
| AR23 | .52 | | -0.68 | VR16 | .50 | | -0.63 |
| AR24 | .54 | | -0.79 | VR17 | | .52 | -1.68 |
| AR25 | .62 | | -1.19 | VR18 | .37 | | -0.96 |
| NR1 | | .60 | -1.06 | ~~*VR19*~~ | | | |
| NR2 | | .81 | -0.65 | VR20 | .25 | | -0.81 |
| NR3 | | .63 | -0.77 | VR21 | | .40 | -1.37 |
| NR4 | | .70 | -1.33 | VR22 | .58 | | -0.78 |
| NR5 | .80 | | -1.23 | VR23 | .63 | | -0.34 |
| NR6 | .58 | | -1.22 | VR24 | | .47 | -1.07 |
| NR7 | | .62 | -1.00 | VR25 | .37 | | -0.88 |
| NR8 | | .68 | -1.30 | VR26 | | .32 | 0.12 |
| NR9 | | .47 | -0.77 | VR27 | .40 | | -0.64 |
| NR10 | | .55 | -0.90 | ~~*VR28*~~ | | | |
| NR11 | | .70 | -1.04 | VR29 | | .36 | -1.34 |
| NR12 | | .74 | -1.24 | VR30 | | .46 | -0.71 |
| NR13 | | .61 | -1.00 | VR31 | | .56 | -1.08 |
| NR14 | | .59 | -1.12 | VR32 | .37 | | 0.26 |
| NR15 | | .69 | -0.81 | VR33 | | .55 | -0.89 |
| NR16 | | .80 | -0.91 | VR34 | | .44 | -1.37 |
| NR17 | | .71 | -0.91 | VR35 | | .50 | -0.16 |
| NR18 | | .74 | -0.82 | | | | |

Measurement invariance across sex was then tested using Multiple Groups Confirmatory Factor Analysis. The two-factor model showed an acceptable fit in the female ($\chi^2_{(3001)}$ = 3229, *p* < .001 [normed chi-square 1.08]; CFI = .98; RMSEA = .012) and male ($\chi^2_{(3001)}$ = 3714, *p* < .001 [normed chi-square 1.24]; CFI = .97; RMSEA = .014) groups, indicating that testing measurement invariance could proceed.

Although several items violated measurement invariance across sex, the RMSEA and CFI values indicate that the decrements in fit were not large (see Table 7). Furthermore, few items were problematic, indicating that the GRT2 is largely measurement invariant across sex at the item level. Only items from the verbal reasoning subtest violated measurement invariance, and many of these items were located towards the end of this subtest. This is unlikely to be related to any item difficulty effects, given that item difficulty does not increase over the subtest. After freeing the loadings and thresholds of Verbal Reasoning items 11, 29, 30, 31, 34 and 35, the factor structure of the GRT2 was found to be partially measurement invariant across sex. The latent mean for males was higher for both Gf (*z* = 7.65, *p* < .001, *d* = .44) and Gc (*z* = 7.87, *p* < .001, *d* = .45), indicating some bias at the latent factor level despite little bias in the items. This supports the ESEM results suggesting that sex influenced the mean of Factors I and II, with a male advantage on both factors. There were fewer items violating measurement invariance than there were items identified as showing a sex difference in the ESEM analysis; however, all items violating measurement invariance in Sample 2 were also found to demonstrate DIF in Sample 1. Additionally, there did appear to be any relationship between item type (series, analogies, etc.) and DIF across sex.

Table 7

*Measurement invariance statistics*

| | $\chi^2$ | df | RMSEA | CFI | Δ $\chi^2$ | Δ df | Δ *p* |
|---|---|---|---|---|---|---|---|
| Baseline | 6873 | 6002 | .013 | .975 | | | |
| Model 2 | 7002 | 6077 | .013 | .973 | 143 | 75 | <.001 |
| Model 2.2[1] | 6984 | 6076 | .013 | .974 | 130 | 74 | <.001 |
| Model 2.3[2] | 6965 | 6075 | .013 | .974 | 117 | 73 | <.001 |
| Model 2.4[3] | 6953 | 6074 | .013 | .974 | 108 | 72 | .004 |
| Model 2.5[4] | 6942 | 6073 | .013 | .975 | 100 | 71 | .014 |
| Model 2.6[5] | 6931 | 6072 | .013 | .975 | 92 | 70 | .041 |
| Model 2.7[6] | 6923 | 6071 | .013 | .975 | 85 | 69 | .088 |

[1]VR29; [2]VR30; [3]VR35; [4]VR31; [5]VR11; [6]VR34

## Multitrait-Multimethod Analysis

Multitrait-multimethod (MTMM) modeling is typically concerned with evaluating the construct validity of a measure, including evaluating the convergent validity of different method factors across each trait, discriminant validity of trait factors, and discriminant validity of method factors (method effects). Analysis can be performed using either the average of all items belonging to each combination of trait and method (trait-method unit) as the manifest variables from which trait and method factors are estimated, or higher-order CFA can be used to first estimate a latent variable for each trait-method unit and then, from these first-order factors, trait and method factors can be modeled at the second order (Marsh & Hocevar, 1988). The second method is preferable as it allows modeling of error and assessment of whether the hypothesised trait-method structure accurately reflects the data (Marsh & Hocevar, 1988).

Additionally, with both of these methods, analysis of a series of nested models, as per Widaman (1985) is the preferred method of assessing construct validity, because it allows explicit tests of the presence of convergent validity and discriminant validity through model comparison. However, estimation problems are often encountered with these models, particularly with the baseline correlated trait-correlated method model, and when the number of traits or methods is less than three. The correlated uniqueness model (Kenny, 1976; Marsh, 1989) is less prone to estimation problems. In this model, only trait factors are modeled, while method factors are conceptualised as correlated residuals amongst those trait-method units sharing method variance. This model can be applied as a second-order CFA model (Eid, Lischetzke, Nussbek & Trierweiler, 2003; see Figure 3).

Two conceptually different types of MTMM model were considered in the present study. Firstly, the construct validity models: a two-trait-factor model where the Gf and Gc factors identified in the present study as representing the structure of the GRT2 were modeled as trait factors, and the presence of method effects due to 1) item stimuli (abstract, numeric, verbal), or 2) item type (analogies, series, classification, maths, vocabulary) was assessed. Secondly, a three-trait-factor model that expressed item stimuli (abstract, numeric and verbal reasoning) as a latent "trait" factor, and specified item type (analogies, series etc.) as a "method" factor was proposed to examine the relative contribution of stimulus and item type. Estimation problems were encountered when fitting the first-order CFA model representing item stimuli as the "trait" and item type as the "method". This appeared to be due to high correlations between latent variables representing trait-method units and indicated that the hypothesised trait-method structure was inappropriate for the data. As such this model was not investigated further.

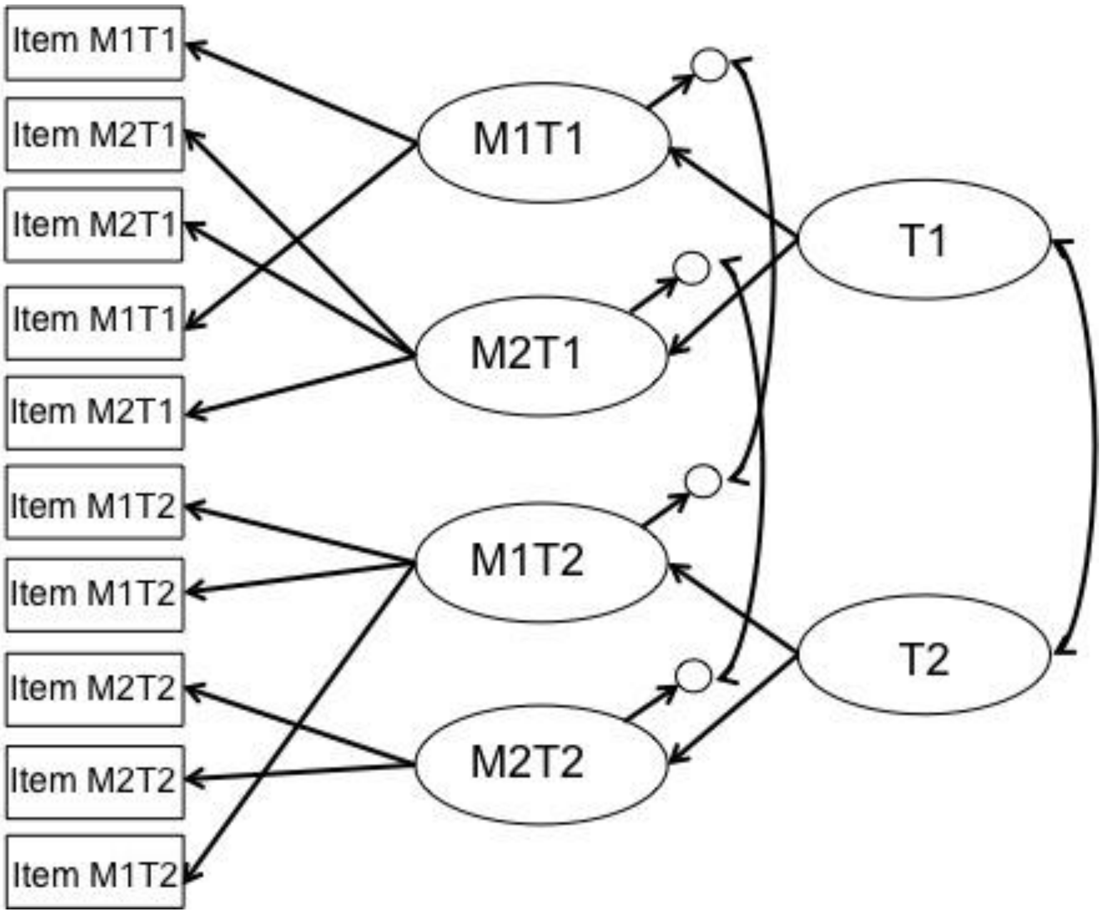


*Figure 3.* Second-order correlated uniqueness model. M1T1 = Method 1, Trait 1; M2T1 = Method 2, Trait 1; M1T2 = Method 1, Trait 2; M2T2 = Method 2, Trait 2.

**Construct validity models.** Firstly, the first-order CFA model was estimated to determine whether the hypothesised trait-method structure was appropriate for the item stimuli as a method factor model. This model demonstrated acceptable fit to the data ($\chi^2_{(2915)}$ = 3679, *p* < .001 [normed chi-square = 1.26]; CFI = .98; RMSEA = .012).

Estimation problems were encountered with this model when a second-order correlated trait-correlated method model was fit, likely due to the presence of only two trait factors, and also only two method factors. The third method factor (figural) could not be modeled, because there was only one figural item loading on Gc.

Therefore the second-order correlated uniqueness model as described in Eid et al. (2003) was used. This model demonstrated good fit to the data ($\chi^2_{(2917)}$ = 3681, *p* < .001 [normed chi-square = 1.26]; CFI = .98; RMSEA = .012). All trait-method units representing both ability factors loaded significantly, and highly, on their respective factors, with standardised estimates ranging from .83 to .96. However, the two traits were very highly correlated, at .95. Although the residual method correlation for numeric stimuli did not account for any variance, the residual method correlation for verbal stimuli did (residual correlation = .40, *p* < .001).

The first-order CFA model for the item type as a method factor model could not be fit to the data due to high correlations between several trait-method units. This indicated that the hypothesised trait-method structure was not appropriate, and this model was not investigated further.

## Discussion

The aim of this study was to investigate the psychometric properties of the GRT2, in terms of its factor structure, sex and age differences, and construct validity. Both EFA and CFA supported a two-factor (Gf-Gc) structure, contrary to previous research conducted in the British population (Moutafi et al., 2005; Furnham et al., 2005), but in line with the theoretical explanations and construction of the test.

Consistent with theory, abstract reasoning items tended to load onto the Gf factor and numerical reasoning items tended to load onto the Gc factor, while verbal reasoning items showed a more complex pattern of loadings, with some items loading on Gf and others on Gc.

Several items showed no demonstrable loading on either factor. These items were VR2, VR3, VR8, VR19 and VR28. Item VR19 has previously been identified as problematic within the Australian population, with individuals scoring correctly on this item tending to score incorrectly on other verbal reasoning items. In fact, the correlation between this item and scores on the verbal reasoning subtest as a whole is only .09 (*r* = .03 for total GRT2 score). This is perhaps due to cultural differences in the use of the stimulus word. It is unclear what the cause of the other mis-fitting items might be; answers to these other items were not so inconsistent with the remainder of the verbal reasoning items, nor with the test as a whole. However, it is conceivable that there may also be cultural differences in the subtleties of the words used in these items, which cause them to be somewhat inaccurate measures of the required constructs within the Australian population.

Sex differences favouring males in the mean factor scores were indicated from both the ESEM and MGCFA analysis. These results support the raw score differences found in all subtests (although this raw score difference was only significant in the case of numerical reasoning). It is interesting to note that there is little sex difference in factor scores under the age of 20 (see Appendix II). Given that it is generally accepted that there is no sex difference in general intelligence (Camarata & Woodcock, 2006; Keith et al., 2011), it may be the case that after secondary school males have more exposure to education and careers associated with the types of abilities assessed in this measure, particularly numerical and abstract reasoning. For example, Valla and Ceci

(2014) suggest that despite increasingly superior results of females in primary and secondary education, they are under-represented in science, technology, engineering and mathematics fields. Training in such fields could conceivably increase familiarity with reasoning using abstract figures and numerical stimuli. Supporting this, more males than females answered all items, indicating that males may have felt more confident in answering these items.

A small number of items were identified as demonstrating DIF across sex in the MGCFA model and ESEM analysis; however, the differences in model fit were not large. While the GRT2 can be largely considered measurement invariant across sex, caution should be used in interpreting the scores of some of the items identified as displaying sex DIF, which are mostly found in the verbal reasoning subtest.

A small number of items were also identified as demonstrating DIF across age in the ESEM analysis. These effects varied in direction. Furthermore, the effect of age at the factor level was only significant for the Gc factor (Factor II). It is interesting that only one abstract reasoning item showed DIF according to age, and there was no significant effect of age on the Gf factor. This is interesting because abstract reasoning raw scores, the largest proportion of items loading on the Gf factor, show a significant negative correlation with age according to the Australian norms. An age difference in Gf would follow from theory, as well as previous research (Salthouse, 2016). It may be that the age range considered in this analysis was not sufficiently broad to identify this effect. Only 12% of respondents included in the current analysis were over 44 years old, which is when this effect has been identified in raw scores (GeneSys Australia, 2012) and, furthermore, only 1% were over 55 years of age. Age did, however, have a significant effect on the Gc factor. The positive influence of age on Gc is consistent with the literature (Salthouse, 2016).

Inspection of the initial CFA results indicated that item type had some influence over the factor loadings of the items belonging to the verbal reasoning subtest. However, item type did not appear to have such an effect with regard to the abstract and numerical reasoning subtests, nor with regard to sex differences identified in individual items. It appears that item type, at least in the case of abstract and numerical reasoning, does not have as large an influence on what is measured as does the item stimulus. Supporting this, there was no evidence for item type as a method factor in the MTMM models, indicating that the specific item type does not account for substantial variance in test items after accounting for ability (Gf/Gc) or stimulus.

Examination of construct validity using MTMM modeling indicated that the GRT2 two-factor structure has good convergent validity when considering item stimuli as a potential method factor. However, there was some evidence against discriminant validity of the abilities when accounting for stimulus type as a method factor. Unfortunately, under a correlated uniqueness model, estimates of trait variance are often inflated, and this can lead to underestimates of discriminant validity (Lance, Noble, & Scullen, 2002), in part due to the assumption of orthogonal method factors. This could therefore be a result of the model applied, particularly given the relatively high correlation between the Gf and Gc factors in both the EFA and CFA analyses.

Overall, the theoretical factor structure of the GRT2 was supported in the Australian population, indicating that the GRT2 likely does measure Gf and Gc. The measure was shown to be largely measurement invariant across sex, although there was evidence for a male advantage at the latent level. Additionally, several verbal reasoning items were identified that may be problematic for use in the Australian population. These last two points should be kept in mind when evaluating and interpreting scores on the GRT2 in the Australian context.

# Chapter 7: General Discussion and Research Conclusions

The research presented in this thesis examined the role of visuospatial ability (Gv) in the Raven’s Progressive Matrices (RPM), and how this might relate to sex differences observed on this test. Additionally, the relationship between Gv and fluid ability (Gf), particularly inductive reasoning, was of interest. The influence of the use of figural stimuli for measures of inductive reasoning was explored in terms of its relationship to sex differences. Four studies were presented to investigate these issues. This chapter presents a brief summary of the conclusions related to each paper, followed by a discussion of the significance and implications of this research. The chapter ends with a discussion of limitations and suggestions for future directions.

## Research Conclusions: Paper 1

Paper 1 assessed the dimensionality of three different versions of the APM using confirmatory factor analysis and Rasch modeling. Sex differences were also considered.

From this paper it can be concluded that the APM is largely a unidimensional measure, for all versions considered and across sex. Therefore, if Gv is involved in performance, it appears to be involved to a similar extent in all items. The one possible exception to this was the difficulty models, which did show reasonable fit to the data. It has previously been suggested that models such as these may reflect qualitative as well as quantitative changes (Vigneau & Bors, 2008) and that two-dimensional models could be related to a position effect (Schweizer et al., 2009). However this could not be ascertained from the results of this study.

It can also be concluded that the solution taxonomies identified in previous literature do not account for the factor structure of the test. While previous research had reported this, it had not been investigated in males and females separately.

Finally, the APM was found to be largely invariant across sex, supporting the fact that it does operate in the same way across males and females, and suggesting that a sex difference in the latent structure of the APM across sex could not account for previous contradictory findings. A male advantage on the latent APM factor was found, suggesting that the sex difference previously reported on this test is not restricted to manifest scores.

## Research Conclusions: Paper 2

Paper 2 used structural equation modeling to examine the unique contribution of Gv abilities to performance on the RPM tests over and above Gf abilities. Additionally, sex differences in the relationships between Gv, Gf and the RPM were investigated.

From this paper, it can be concluded that, although Gf explains a large portion of the variance in both latent APM and SPM, Gv has a non-trivial influence on explaining the residual variance. This study has provided evidence that Gv abilities are involved in the RPM tests. The strongest contender for a narrow Gv ability to be implicated was flexibility of closure; however this could only be confirmed in one sample (Sample 3 – SPM).

There was little evidence for a different relationship between Gf, Gv and latent APM between males and females (APM Sample 1), indicating that Gf and Gv contribute in a similar way to APM performance across both sexes. Multiple Indicator Multiple Causes modeling indicated that accounting for the sex difference in Gv could explain the sex difference in latent APM.

## Research Conclusions: Paper 3

Paper 3 used meta-analytic techniques to summarise research concerning sex differences in manifest scores on measures of inductive reasoning. The item stimulus and item type used in these measures were investigated as moderators of the sex difference.

From this paper it can be concluded that there is substantial variance in the sex differences identified on measures of inductive reasoning, particularly across different types of measures, but also within individual measures, making it difficult to propose strong conclusions regarding the magnitude of any consistent sex differences.

However, there was a clear difference between figural and alphabetic measures considered. The RPM tests, and other figural inductive reasoning tests, tended to show a small-to-moderate male advantage. On the other hand, letter series tests showed a small female advantage. Moderator analysis indicated that figural stimuli may account for some of the variance in the magnitude of Hedges' *g* across the different measures of inductive reasoning.

## Research Conclusions: Paper 4

Paper 4 used exploratory and confirmatory structural equation modeling to examine the factor structure, sex differences, and role of item stimulus and type in the General Reasoning Test 2 (GRT2). Support was again found for a male advantage on abstract (figural) reasoning in Paper 4. However, the male advantage was present not only on figural reasoning, but on both the Gf and Gc factors identified. It is possible that this occurred in the case of Gf because of the use of figural stimuli and in the case of Gc because of the numerical reasoning items. Raw scores showed a greater sex difference on the numerical and abstract reasoning subtests which composed the majority of items included in Gc and Gf, respectively, and less difference between the sexes on the verbal reasoning subtest which was split across Gf and Gc.

The results of the multitrait-multimethod analysis indicated that the stimulus format of the items did not have a significant influence on the particular ability measured, with little evidence of method effects. Furthermore, generally speaking, the type of question had no consistent impact on the ability measured, although it did appear to influence the factor loadings of the verbal reasoning items.

## General Discussion: Significance and Implications

The findings from this research have implications for the use of the RPM tests specifically, and also for the use of other measures of inductive reasoning. Further evidence was found for the unidimensionality of the Advanced RPM tests, but both the Standard and Advanced versions were found to involve Gv. Additionally, the APM demonstrated a male advantage in both latent means and manifest scores. Figural format of the tests was proposed as a moderator of this advantage. This has implications for the interpretation of sex differences on inductive reasoning measures more broadly as well as our understanding of intelligence theory and how inductive reasoning fits into the currently accepted framework of CHC theory.

### RPM as a measure of Gf

One aim of the present research was to investigate the contribution of Gv to performance on the RPM tests, including examining evidence for multidimensionality within these tests. Results from Paper 1 support the notion that all APM items measure the same construct, for both males and females; this had not been previously established for the Advanced version of the tests. Confirmation of measurement invariance across sex further indicates that a difference in construct measured across sex cannot account for previous contradictory results regarding the dimensionality of the APM tests. Results from Paper 1 support previous research findings that the analytic-visuospatial distinction may not be responsible for any reported findings of multidimensionality (Vigneau & Bors, 2008).

What remains then, is the question of what the “unidimensional” construct measured by the APM may be. The present research suggests the answer to this question may not simply be Gf or inductive reasoning but, instead, a composite of Gf and Gv equally represented across all items included in the test. This is a finding that may threaten the construct validity of the RPM.

## Sex Differences, Gv and the RPM

A male advantage on the RPM was consistently found in this research, consistent with Lynn and Irwing (2004b) and Irwing and Lynn’s (2005) meta-analyses. However, the issue of the male advantage is potentially problematic to the notion that the RPM tests provide an exceptionally “pure” measure of *g*, or of Gf. Although it is debated, research concerning sex differences in latent general intelligence factors has consistently reported no or little difference (Calvin, Fernandes, Smith, Visscher, & Deary, 2010; Camarata & Woodcock, 2006; Keith, Reynolds, Roberts, Winter, & Austin, 2011; Mackintosh, 1996), as has research concerning sex differences in latent inductive reasoning (Arendasy & Sommer, 2012) and latent Gf (Keith, Reynolds, Patel, & Ridley, 2008; Lakin & Gambrell, 2014; Reynolds, Keith, Ridley, & Patel, 2008).

Therefore, either the RPM tests are not an exceptionally pure measure of *g* or Gf, or there is test-specific variance causing the observed sex difference, or both.

Results from Paper 2 also indicated that the RPM tests may not be an exceptional measure of Gf, to the extent that Gv explains a substantial amount of residual variance in latent RPM after accounting for Gf. Furthermore, results indicated that the involvement of Gv could account for the observed sex difference in the APM, indicating that the Gf measured from the RPM tests is at least somewhat contaminated by Gv.

## RPM and Academic Achievement

General intelligence is well established as one of the strongest predictors of academic achievement, predicting around 25% of grades (Neisser et al., 1996). This is not surprising given the historical context of the development of intelligence testing. More recent meta-analyses have found that intelligence continues to be an excellent predictor of academic achievement. For example, Poropat (2009) found a correlation of *r* = .25 between intelligence and academic performance, reduced to *r* = .23 in university samples, while Roth et al. (2015) found a higher correlation of *r* = .54 between intelligence and academic performance in school students. Given the high correlation between Gf and *g*, fluid intelligence in particular could be considered an excellent predictor of academic achievement. Consistent with this suggestion, Postlethwaite (2011) reported a correlation of *r* = .26 for academic achievement and Gf overall, and a slightly lower correlation of *r* = .22 among university samples.

However, the present thesis brings into question the notion of the APM tests as a particularly good measure of general intelligence or even Gf. Therefore, in order to test the predictive validity of the APM for academic achievement, data concerning a sample of *N* = 881 third year psychology undergraduates from the University of Adelaide was analysed. These data had been collected over several years, and included total APM score (12-item short form; Bors & Stokes, 1998), verbal intelligence (total Spot-the-Word score; Baddeley, Emslie, & Nimmo-Smith, 1993), conscientiousness (as measured by the OCEANIC), average course grade across all third year Psychology courses, age, sex, and Tertiary Entrance Rank (TER; this is the score of academic achievement awarded to high school graduates, expressed as a percentile rank measure and used for application to university courses in Australia).

Table 7.1 presents the descriptive statistics for these measures. From Table 7.2, it can be seen that, although APM is significantly correlated with grade, it is not a particularly strong relationship. Multiple regression was used to investigate the incremental validity of APM in explaining course grades after accounting for sex, age, TER, conscientiousness and Gc. Non-normality of residuals was identified in this dataset but transformations did not improve this and, given the large sample size, regression analysis proceeded. This analysis found that APM scores explained only an additional 2% of course grades (Table 7.3). Additionally, relative importance regression (calculated using the R package relaimpo [Gromping, 2006]) indicated that APM scores accounted for relatively little of the 19% explained variance in grades, and much less than either TER or C.

It could be argued that TER and grades will share substantial overlap, given that they are both measures of academic achievement. In fact, Soares, Lemos, Primi, and Almeida (2015) found that reasoning ability no longer accounted for academic achievement once variance in prior academic achievement was accounted for. Therefore, a model excluding TER was also tested (Model 4). When TER was excluded, APM still only accounted for 3% of the variance in grades (as seen from relative importance values indicating APM explains 25% of the 13% explained variance in Model 4).

These results suggest that, while the APM does predict academic achievement, it is not a particularly remarkable predictor. Of course, the correlation between Gf and academic achievement is attenuated by range restriction in university samples. Chamorro-Premuzic, Quiroga, and Colom (2009), for example, reported a correlation of *r* = .17 between Gf and university grades. This finding suggests that perhaps Gf in general may not be a particularly good predictor of grades in University samples (however, see the following section for further comments on this result).

Although Chamorro-Premuzic et al. (2009) reported a similar correlation between Gf and grades as found for the present data, previous research concerning university samples using only the APM to measure Gf has found slightly higher correlations with academic achievement than the present results. Thus, Chamorro-Premuzic and Arteche (2008) reported a correlation of *r* = .23 between general academic performance and a timed version of the APM among university students. Rohde and Thompson (2007) reported a correlation of *r* = .43 between the full APM and SAT scores and *r* = .31 between APM and GPA.

Interestingly, APM was more highly correlated with SAT Maths scores (*r* = .45) and overall SAT scores (*r* = .43) than SAT Verbal scores (*r* = .30) in Rohde and Thompson’s (2007) study. Similarly, in samples of high school students, the RPM has been shown to correlate more highly with mathematics subjects than language subjects (Pind, Gunnarsdottir, & Johannesson, 2003; Rindermann & Neubauer, 2004). Gv has been found to be important in STEM achievement (Wai, Lubinski, & Benbow, 2009), while visuospatial training may improve performance in STEM areas (Sieff & Uttal, 2015). Therefore, it may be that the Gv component of the APM results in stronger correlations with such academic areas, and stronger correlations with overall academic achievement variables that may incorporate results from such subjects. The present analysis concerns grades in Psychology courses, which may not show as strong a relationship with Gv. This may explain the somewhat lower correlation in the present case. This is also consistent with the finding that the RPM involve Gv.

Table 7.1

*Means and standard deviations*

| | Mean | SD | N | Range |
|---|---|---|---|---|
| Age (yrs) | 23.66 | 6.47 | 881 | 18.3 – 67 |
| TER | 86.46 | 9.82 | 690 | 46.2 – 99.95 |
| APM | 7.59 | 2.71 | 880 | 0 – 12 |
| C | 37.89 | 7.30 | 879 | 9 – 54 |
| STW | 24.53 | 3.28 | 877 | 11 – 30 |
| Grade | 69.12 | 11.64 | 881 | 23 – 96.5 |

*Note.* TER = Tertiary Entrance Rank; APM = Advanced Progressive Matrices; C = Conscientiousness; STW = Spot-the-word test; Grade = average grade across all third year Psychology courses.

Table 7.2

*Correlations*

| | Age | TER | APM | C | STW |
|---|---|---|---|---|---|
| TER | -.05 | | | | |
| APM | .08* | .07 | | | |
| C | .10*** | .09* | -.06 | | |
| STW | .21*** | .19*** | .23*** | .02 | |
| Grade | -.01 | .33*** | .18*** | .26*** | .14*** |

*Note.* TER = Tertiary Entrance Rank; APM = Advanced Progressive Matrices; C = Conscientiousness; STW = Spot-the-word test; Grade = average grade across all third year Psychology courses. *** *p* < .001; * *p* < .01

Table 7.3

*Regression models*

| | B | SE B | *β* | *t* | RI | $R^2$ | $\Delta R^2$ |
|---|---|---|---|---|---|---|---|
| Model 1 | | | | | | .02 | |
| Sex | -2.99 | 1.02 | -.11** | -2.94 | .70 | | |
| Age | 0.20 | 0.10 | .08* | 2.03 | .30 | | |
| Model 2 | | | | | | .17 | .15*** |
| Sex | -1.54 | 0.98 | -.06 | -1.58 | .04 | | |
| Age | 0.16 | 0.09 | .06 | 1.67 | .02 | | |
| TER | 0.32 | 0.04 | .29*** | 7.98 | .55 | | |
| STW | 0.24 | 0.13 | .07 | 1.79 | .05 | | |
| C | 0.33 | 0.5 | .05*** | 6.22 | .33 | | |
| Model 3 | | | | | | .19 | .02*** |
| Sex | -1.84 | 0.96 | -.07 | -1.90 | .04 | | |
| Age | 0.13 | 0.96 | .05 | 1.41 | .02 | | |
| TER | 0.32 | 0.04 | .28*** | 7.86 | .46 | | |
| STW | 0.11 | 0.13 | .03 | 0.83 | .03 | | |
| C | 0.35 | 0.05 | .23*** | 6.66 | .31 | | |
| APM | 0.67 | 0.15 | .16*** | 4.53 | .13 | | |
| Model 4 | | | | | | .13 | |
| Sex | -2.81 | 0.90 | -.10** | -3.12 | .08 | | |
| Age | -0.11 | 0.06 | -.06 | 0.06 | .01 | | |
| STW | 0.41 | 0.12 | .12*** | 3.49 | .13 | | |
| C | 0.41 | 0.05 | .26*** | 8.04 | .53 | | |
| APM | 0.77 | 0.14 | .18*** | 5.49 | .25 | | |

*Note*. RI = Relative Importance. * $p < .05$; ** $p < .01$; *** $p < .001$.

### Alternative Explanations

Despite findings in the present thesis indicating that the RPM tests do involve Gv to a substantial extent, and that sex differences exist in performance, the large body of literature reporting high loadings of the RPM on a *g* factor requires an explanation. This presents a paradox; on one hand, the RPM tests have been found to be an excellent measure of *g* and Gf, but on the other, present findings contradict the notion that they are solely a measure of Gf.

The position effect in Gf measures has been widely researched by Schweizer and colleagues (Ren, Goldhammer, Moosbrugger, & Schweizer, 2012; Ren, Schweizer, Wang, & Xu, 2015; Ren, Wang, Altmeyer, & Schweizer, 2014; Schweizer, 2012; Schweizer, Reiss, Schreiner, & Altmeyer, 2012; Schweizer et al., 2009; Schweizer, Troche, & Rammsayer, 2011). This research may help to explain both the contradictory results regarding the APM as a predictor of academic achievement and the APM as an exceptional measure of Gf.

Regarding the prediction of academic achievement, this research suggests that it is the learning component, or position-specific variance, of Gf measures that may be more important in predicting academic achievement than the ability-specific variance (Ren et al., 2015). A decrease of the position effect due to the use of a short form may therefore attenuate the relationship between the APM and academic achievement. Interestingly, Chamorro-Premuzic et al. (2009) used screening versions (i.e. shorter forms) of two of their Gf measures (APM and DAT-AR), and their study showed a similar pattern of correlations between Gf and university marks and Gf and university entrance scores to that in the present data, which came from a short form of the APM.

In terms of the broader issue of the RPM as an excellent measure of Gf, that performance variance on Gf measures may be partitioned into position-specific and ability-specific variance raises the conceptual question of what this means for Gf abilities; is Gf related to only one of these components, or does it necessarily require both? If what we mean by Gf is captured by the position-specific variance, then perhaps a short form of the APM is not adequate. However, if Gf is to be interpreted as ability-specific variance only, then a short form of APM should suffice. Schweizer et al. (2011) reported that the ability-specific portion of Gf, measured by Horn's numerical reasoning scale, is more highly correlated with general intelligence than the position-specific component. This indicates that if we subscribe to the notion that Gf is equivalent to *g* (e.g., Gustafsson, 1984), then the decrease in position-specific variance in a short form is not detrimental to the measurement of Gf. However, if Gf is not isomorphic with *g*, it could be that the position effect is an important component of Gf.

Research by Bui and Birney (2014) has additionally suggested that learning is not a necessary component of performance on the RPM tests, but that learning does occur during testing and can influence performance. Specifically, actualised learning, where the individual is not just exposed to the item rules, but successfully solves the item, accounts for some individual variance in RPM performance. This provides additional support for the idea that reducing the amount of learning may influence exactly what the test measures.

Ren et al.'s (2012) research suggests that ability-specific variance, the relative proportion of which may increase with the use of a short form, is more highly related to perceptual attention processes than the position-specific variance. This could account for some of the shared variance between the RPM and Gv measures. The contrast between results produced in Paper 2 as compared to those of Schweizer et al. (2007) does raise the question of whether the construct measured by a short form of the APM is the same as that measured by the full form, and whether the position effect could account for the difference. Bors and Stokes' (1998) original study indicated that their short form correlated highly with the full form (*r* = .90) and the correlation between Shipley Abstraction scores and the short form was not significantly different from the correlation between Shipley Abstraction and the full form. However, to our knowledge, the effect of learning on the short form as compared to the full form has not been extensively studied.

Although research concerning the position effect provides a potentially tenable argument for the contradictory findings, a problem is encountered in that Paper 1 demonstrated the APM to be unidimensional. This unidimensionality indicates that each item measures the same thing, and to the same extent, whatever that might be. This would indicate that even though the absolute level of position-specificity is reduced in using a short form, the relative level is unlikely to be. On the other hand, the position effect may help explain a "qualitative" difference between beginning and end items of the APM, causing the acceptable fit of difficulty models as identified in Paper 1 and by Vigneau and Bors (2008).

Finally, there may be some difference in the construct measured depending on whether a timed or untimed version is administered. However, results from Paper 2 included analysis pertinent to both a timed (APM Sample 2 and Sample 3) and untimed (APM Sample 1) administration of the test. Although the contribution of Gv was slightly lower in APM Sample 2, it was unclear whether this was due to timed administration, or the use of different measures to represent the alternate constructs. Regardless, Gv was still relevant to performance on both timed and untimed administration of the RPM, indicating that the contribution of Gv to RPM performance is not unique to either timed or untimed administration.

## Implications for Intelligence Theory and Measurement

The possibility that the RPM tests involve Gv necessitates a consideration of the relationship between Gf, the ability purported to be measured by the RPM, and Gv. This relationship is complex and not fully understood; and the fact that many of the most widely used measures of Gf, and specifically inductive reasoning, utilise figural stimuli may confuse the relationship between these two abilities. Paper 2 demonstrated that Gv is involved in performance on the RPM tests, despite these tests being considered one of the best measures of Gf. These findings have implications for the role of a “content facet” and the effect that might have on the measurement of Gf.

The results from Paper 2 suggest that the use of figural stimuli may recruit involvement of Gv abilities into performance. Similarly, findings from Paper 3 indicate that the stimulus used to measure inductive reasoning influences sex differences found in manifest scores. However, findings from Paper 4 indicate that stimulus may not substantially influence the underlying latent ability measured, nor be related to sex differences on individual items. In the case of Paper 4, this may have been due to the fact that all items were not strictly measures of Gf: several verbal reasoning items were vocabulary, which can more appropriately be considered Gc, while several numerical reasoning items could more appropriately be considered measures of quantitative ability (Gq). It may be that the content facet becomes more important when all items measure the same process, for example, inductive reasoning.

## Measurement of Inductive Reasoning

The results of Paper 3 are problematic for the measurement of inductive reasoning. Although sex differences per se cannot determine what a test measures, the fact that such a wide range was found in the observed effect size of the sex difference, particularly comparing letter series (PMA-R) and figural series (DAT-AR) measures, indicates that there is some difference in what is being measured. Additional support for this comes from aging research; letter series measures have also been shown to behave differently from other inductive reasoning measures in older populations. It is a well-accepted finding that Gf declines with older age but research has identified some differences in the relationship between Gf and aging depending on the measure used. Thus, Salthouse (2005a, 2005b) and Habeck et al. (2015) have presented results indicating that matrix reasoning declines at a faster and more pronounced rate than does solution of series completion items, as measured by the Shipley Abstraction test and letter series items. Additionally, Burns, Burns and Ward (2016) reported that although scores of younger adults on the APM were significantly higher than those of older adults, there was no difference in scores across age groups for the CAB-I (although others do present findings indicating that letter series measures decline at a similar rate; see Klein, Dilchert, Ones, & Dages, 2015).

**Stimulus type, Gf and Gc.** It is certainly the case that if we make the argument that the figural stimuli used in figural matrices tasks such as the RPM recruit the involvement of Gv, the stimulus type used in letter series task may lead to Gc contamination. This could be an explanation for the difference in effect size between alphabetic and figural inductive reasoning measures found in Paper 3. This would additionally indicate that it is inappropriate to use either of these types of test as a single measure of Gf. However, it should be noted that in Paper 3, verbal analogies and verbal classification tests did not show the same magnitude of female advantage as did the PMA-R. Furthermore, sensitivity analysis showed that the effect of verbal versus figural stimuli may not be as robust as the effect of alphabetic versus figural.

If the argument is that involvement of Gc is contributing to the female advantage in the case of the PMA-R, it is unclear why the same effect is not apparent in verbal tests. If anything, verbal analogies would likely demonstrate stronger loadings on a Gc factor than letter series tests; acculturation to the alphabet is much simpler than abstract meanings of words. It therefore seems that it could be something specific about the use of alphabetic stimuli that causes this difference (although it was also unclear, because of availability of only a small number of data points, whether the same magnitude of female advantage found for PMA-R was apparent in other alphabetic series and classification measures).

It is worth noting that, although factor analyses have sometimes found the RPM to load on both Gf and Gv (Rosen, 1995) -- or a combined Gf/Gv factor (Crawford, 1991) -- and verbal analogies or classification tests to load on both Gf and Gc (Johnson & Bouchard, 2005b; Kleitman & Stankov, 2007), a similar phenomenon regarding combined Gf-Gc loadings has not commonly been found for letter series, sets or grouping tests. For example, Rosen (1995) found that the RPM loaded on both *g* (a Gf factor was not included) and Gv, while letter grouping was found to load only on *g*. In Crawford’s (1991) study, letter series loaded onto a Gc factor when a combined Gf/Gv factor was included in the solution, but only on Gf when the Gv ability measures were excluded from analysis. However, although Johnson and Bouchard (2005b) found that the CAB-I loaded only onto a fluid ability factor when a Gf-Gc model was applied to their data, the CAB-I loaded on spatial and number factors (Perceptual and Verbal abilities) when the Verbal-Perceptual-Image Rotation model was applied.

It is also important to note that some research has suggested that some types of verbal reasoning may also involve Gv. Thus, Colom, Contreras, Arend, Garcia Leal, and Santacreu (2004) found that sex differences in a measure of verbal reasoning, the three-term series, were mediated by Gv. Although the three-term series tests are somewhat different from the tests considered in Paper 3, this result does suggest that we cannot assume that, simply because the PMA-R test does not include visuospatial stimuli, Gv processes cannot be engaged, or that Gc necessarily is.

**Stimulus familiarity**. Meo et al.'s (2007) element salience hypothesis proposed two sources of item difficulty in the RPM: object overlap and stimulus familiarity. Paper 2 in this thesis presented support for the involvement of object overlap, finding that flexibility of closure was the strongest narrow Gv predictor of SPM performance. It is possible, however, that stimulus familiarity may become relevant when comparing alphabetic and figural tests. Meo et al. argued that more familiar stimuli are easier to name, which makes them easier to represent mentally; the suggestion being that stimuli that are harder to name are also therefore more taxing on working memory capacity.

Although Meo et al. (2007) found no main effect of item type on accuracy (comparing Raven's matrices items, and other matrices items utilising European letters, overlapped European letters and invented letters), when complexity increased, accuracy decreased the most for Raven's items, which are overlapping and unfamiliar, and the least for European letters, which were not overlapping and not unfamiliar. As the authors argued, this indicates that one moderator of accuracy is the familiarity of the stimulus.

With regard to the results of Paper 3, perhaps it is the case that there are differences between males and females in familiarity with alphabetic and figural stimuli. In any case, differences in scores according to stimulus familiarity would suggest that the stimulus used to measure inductive reasoning can have a non-trivial effect on scores, and therefore may be influential in the particular ability, or combination of abilities, responsible for performance. The results presented in Paper 3 support the notion that the stimuli used does affect measurement in some way, while the results of Paper 2 suggest that figural stimulus may be associated with Gv.

## Sex Differences in Inductive Reasoning

Given the established male advantage on certain tests of Gv, a second aim of the present thesis was to examine whether this has relevance for improved understanding of evidence for sex differences on the RPM, and on other measures of inductive reasoning. In the present body of research, a consistent male advantage was found on the RPM tests, and on other measures of figural inductive reasoning. This is interpreted as an advantage on the RPM and figural reasoning tests specifically, and not as a male advantage in general intelligence. There is substantial evidence against a sex difference in general intelligence at a latent level, and given results indicating the involvement of Gv in RPM performance, it does not appear to make sense to interpret this finding as a male advantage in general intelligence.

Lynn (1999) has proposed that males have higher average intelligence than females, beginning in early adulthood. More recently, the body of research that Lynn and colleagues have used to support this claim has largely been drawn from results on the RPM tests (Irwing & Lynn, 2005; Lynn & Irwing, 2004b), although not always (Colom & Lynn, 2004; Irwing, 2015). Given the present findings that the RPM requires

Gv for successful performance, and that the figural content of inductive reasoning measures moderates the male advantage, the conclusions of Lynn and colleagues are problematic.

## Test-specific versus Content-specific Variance

Results from Paper 3 indicated that the figural content of measures may be related to a larger male advantage. Although previous research has indicated differences between the RPM and other figural matrices tasks (Arendasy & Sommer, 2012) or other figural inductive reasoning tasks (Colom & Garcia-Lopez, 2002; Colom, Stein, et al., 2013), results from Paper 3 have indicated that the sex difference in figural inductive reasoning tasks is not unique to the RPM. The RPM tests show a similar magnitude of sex difference to the WAIS Matrix Reasoning test, and other figural inductive reasoning measures, although the DAT-AR (figural series) showed a slightly higher effect size. When sensitivity analysis was performed in Paper 3, based on proportions of males and females, the DAT-AR effect size did decrease, but it remained similar to that found for the RPM tests. Although meta-analyses and systematic reviews have previously been conducted on the RPM tests (Court, 1983; Irwing & Lynn, 2005; Lynn & Irwing, 2004b), the analyses presented in Paper 3 provide the first synthesis comparing different types of inductive reasoning measures.

The results from Paper 3 demonstrate that the sex difference identified in previous literature, and in Paper 1 and 2, may not be related to test-specific variance, but rather to content-specific variance shared across figural measures of inductive reasoning. The existence of sex differences on the RPM and other figural reasoning tests is contradictory to the observation that these tests measure an ability generally regarded as not demonstrating sex differences at the latent level. However, although the reason for the relationship between the male advantage and figural stimuli is hypothesised to be the involvement of Gv, the results from the meta-analysis were unable to determine this.

## Figural and Alphabetic Inductive Reasoning

The present research suggests that the figural content present in the RPM tests may be at least partially responsible for the male advantage, while the use of alphabetic stimuli may be partially responsible for a female advantage. However, the use of figural stimuli is certainly not the only difference in the tests considered. The relationships between the stimuli in the case of the PMA-R may be of a different complexity to those found in the RPM tests, or even the DAT-AR tests, so that it is difficult to disentangle whether it is the figural content per se, or another factor related to this that accounts for different outcomes. Furthermore, it should be noted that nearly all PMA-R studies included in Paper 3 were from Spanish samples, authored by Roberto Colom. Although papers published by Roberto Colom did not report a significantly different overall effect size as compared to other papers, this could not be tested for the PMA-R specifically. Additionally, other alphabetic tests tended to show smaller and more contradictory effect sizes, although there were insufficient data points to permit strong conclusions. It is therefore somewhat unclear if such a large difference would exist between alphabetic and figural stimuli if the PMA-R was excluded from analysis, or more other types of alphabetic inductive reasoning measures were included.

## Developmental Differences

Results from Paper 4 indicated that the male advantage on Gf, which was largely composed of abstract reasoning items, did not become apparent until after school age. This is in line with Lynn's (1999) developmental theory. All other samples concerned individuals beyond the age of potential developmental effects as it relates to Lynn's theory. However, an alternative explanation of the appearance of a male advantage can be proposed; it may be familiarity with figural-type stimuli, developed by early adulthood, rather than superior general intelligence that causes the male advantage, as discussed in considering the role of stimulus familiarity. Of course, this cannot discount Lynn's developmental theory, but does provide an alternative explanation that accounts for a male advantage on figural measures despite no sex difference in general intelligence.

## Manifest and Latent Differences

It is important to note that the results of Paper 3 concern manifest scores, while the results of Papers 1, 2 and 4 are largely focused on latent differences. Both types of difference are relevant to the use of ability measures because manifest scores are often used when interpreting the results of these measures, while latent differences may be used to examine differences in the underlying construct.

Findings from Papers 1 and 2 report sex differences at the latent level, albeit in a single measure. A latent mean difference favouring males in the Gf factor identified in the GRT2 was also found in Paper 4. Given the results of Paper 3, as well as Papers 1 and 2, the male advantage on the GRT2 Gf factor, which largely consisted of abstract (figural) reasoning items is not surprising. It is interesting to note, however, that this latent male advantage on Gf is contrary to findings indicating no male advantage at the latent Gf level (Keith et al., 2008; Lakin & Gambrell, 2014). However, Keith et al.'s (2008) analysis concerned measures of Gf that, although visual in nature, tend not to be as visually complex as other figural measures considered here, while Lakin and Gambrell's study included a much wider variety of Gf measures. In contrast, the Gf factor from the GRT2 consisted largely of figural items but also included items such as verbal analogies, which also tend to show a male advantage (Halpern, 2012). Surprisingly, a male advantage was also identified on the latent mean of the GRT2 Gc factor. This may be due to the majority of Gc items consisting of numerical reasoning items, which also tend to show a male advantage (Keith et al., 2008).

Overall, the results concerning sex differences suggest that there is a male advantage on figural reasoning at both the manifest and latent level; however the latent difference may be more appropriately thought of as a difference in the more narrow latent variable of figural inductive reasoning, rather than necessarily a difference in Gf.

### Strategy Differences

The notion of strategy differences between males and females on the RPM was not explicitly considered in the present research, but results do provide some suggestions. Measurement invariance of the APM, as found in Paper 1, would suggest that there may not be strategy differences. Although strategy differences are distinct from ability differences and measurement invariance cannot determine how individuals solve a task, it is likely that if different strategies were consistently used across groups, there would be some difference in the relationship of the items to the latent variable.

Furthermore, results from Paper 2 considered whether there was variance in the relationship between Gv and Gf measures and RPM across sex. Results confirmed invariance, indicating that there may not be any consistent differences in strategy used.

However, this conclusion assumes that Gf is correlated with an analytic strategy while Gv is correlated with a visual strategy; and this may not be the case. It is possible that an individual may choose to use a visual strategy on both Gf and Gv tests.

## General Discussion: Limitations and Future Research

### RPM and Gv

#### Flexibility of Closure

Paper 2 presented evidence indicating that Gv is involved in performance on the RPM tests. Results indicated that flexibility of closure may be a key narrow ability involved in performance, in line with previous literature (e.g. Meo et al., 2007). Future research is needed to confirm this and also to investigate the role of flexibility of closure in performance on the APM, which could not be assessed in Paper 2.

#### Components of RPM Performance, Gv and Sex Differences

Among those proposed component processes of performance on the RPM are working memory or goal management (Carpenter, Just, & Shell, 1990; Loesche, Wiley, & Hasselhorn, 2015), placekeeping (Hambrick & Altmann, 2015), strategy use (Gonthier & Thomassin, 2015) and position effects or learning (Ren et al., 2012; Ren et al., 2014; Schweizer et al., 2012; Schweizer et al., 2009). The present research was specifically concerned with whether or not Gv is involved in RPM performance at all. Presents results support the involvement of Gv, which then introduces the question of how Gv might relate to each of these proposed component processes, as well as to the core ability for induction of solution rules.

Research by Schweizer and colleagues (Ren et al., 2012; Ren et al., 2014; Schweizer, 2012; Schweizer et al., 2012; Schweizer et al., 2009) has indicated that the variance on reasoning measures can be partitioned into ability-specific and position-specific variance. It would be interesting to examine where Gv and identified sex differences fit into this type of model. Position-specific variance can be thought to represent working memory and learning, while the ability-specific variance may capture a more “pure” Gf ability. Schweizer and colleagues’ research has indicated that partitioning the variance in such a way can substantially increase the correlations between fluid ability measures when considering the ability-specific portion of variance; however results from Paper 2 indicate that it may affect the construct measured, with the ability-specific portion perhaps more highly related to Gv abilities.

Similarly, the role of working memory, and relatedly, goal management, has been proposed as an important factor in RPM performance (Carpenter et al., 1990; Loesche et al., 2015). Future research should investigate whether the relationship between Gv and RPM is related to these components specifically, perhaps due to some shared relationship with visuospatial working memory. There are certainly indications that visuospatial training may slightly improve performance on the RPM and other figural inductive reasoning tasks, although the evidence is not strong (Colom, Román, et al., 2013; Stephenson & Halpern, 2013).

### Short Forms

Additionally, results indicate that the relationship between short forms of inductive reasoning and long forms should be further investigated. In Paper 2, this was proposed as an explanation for the difference in results to those of Schweizer et al. (2007). Furthermore, Paper 3 identified significant heterogeneity in summary effects for nearly all tests. However, test form was not considered as a moderator, largely because such a variety of different test forms had been used, and there were insufficient data points to account for these different types. It was also the case that several papers had not reported the test form used.

## Sex Differences and Measuring Inductive Reasoning

### Sample Representativeness

A general limitation of the research reported here -- and more broadly speaking with a large amount of the literature on cognitive ability and sex differences -- is the representativeness of the samples used. As discussed in the introduction and literature review, there are many methodological issues with the study of sex differences in cognitive abilities, mainly related to issues of sample representativeness. These should be kept in mind when interpreting the results of the present research. However, the samples used in these studies can be thought of as broadly representative of university students. Furthermore, the data used in Paper 4 was a population sample, and sex differences were also identified in this dataset.

Relatedly, one major limitation of Paper 3 concerns publication bias. Attempts were made to diminish this effect; papers were sought which did not have the aim of investigating sex differences, standardization data was requested and, furthermore, several studies were included which reported sex differences on multiple measures, the reporting of which for each individual test should not be influenced by the significance or lack of significance of the difference. However, it is still the case that these statistics are more likely to be reported when there is a difference than when there is not. Therefore, the effect size estimates of the difference are likely higher than the true effect. However, one of the main aims of this study was to make a comparison among the effect sizes for different types of measure; this is less likely to have been influenced by publication bias, particularly given that several studies reported results for multiple measures. Despite this, future research should confirm findings related to sex differences and stimuli as a moderator of these differences in population-representative samples.

## Letter Series Tests

Paper 3 presented findings indicating that the PMA-R may have unique characteristics when compared to other measures of inductive reasoning. The data collected for this paper were, however, insufficient to permit assessing whether these differences extended to other letter series tests. Although sex differences in the RPM tests have been extensively studied, less attention has been paid to series completion tasks, particularly letter series such as the PMA-R. Similarly, although studies have investigated the cognitive processes required in the solution of figural matrices tasks (Arendasy & Sommer, 2012; Kunda, McGreggor, & Goel, 2013; Primi, 2001), fewer have investigated the processes required in letter and number series completion (Liang, Jia, Taatgen, Zhong, & Li, 2014; Stankov & Cregan, 1993), and particularly how these test types may differ.

Future research should investigate in more depth the PMA-R and other letter series measures to determine what characteristics of these tests might cause a female advantage. This could help in further disentangling the relationship between inductive reasoning, Gv abilities and sex differences.

## Experimental Investigations

All studies included in the present thesis were correlational. These studies were useful for the purposes of investigating the research questions, however they were unable to tell us anything about the casual mechanisms behind the findings. Experimental investigations may be useful in furthering understanding of some of the issues presented here. For example, it would be useful to create equivalent measures of inductive reasoning utilising all possible stimuli types (Verbal, Figural, Alphabetic, Numeric) and item types (Matrices, Series, Classification, etc.) identified in the present studies. This would allow comparison among the degree of difficulty and sex differences in these different measures. Although comparisons among different types of measures were made in Paper 3, there were several stimulus-type categories that do not currently exist in commonly used tests (e.g. letter matrices, letter analogies) and, furthermore, there are differences other than the stimuli used among the items, such as the type of rules to be induced. If equivalent forms could be created, it may help to disentangle the processes behind the test differences.

Studies have been conducted to investigate the influence of familiarity versus unfamiliarity of the stimulus used in matrices tasks (Meo et al., 2007; M. J. Roberts et al., 2000), and this has included examining matrices consisting of figural stimuli and alphabetic stimuli. More research in this vein may be helpful to further understanding of the distinctions between measures using the same item type but different stimuli.

### Strategy Use

Paper 2 provides some information regarding speculation on differences in strategies used to solve the RPM items, and how this might impact on sex differences. However, strategy use was not directly assessed. Investigation of strategy as a mediator of the relationship between Gv, RPM performance and sex differences could be useful to further understanding of the causal mechanisms of this relationship.

## Final Comments

Although understanding of intelligence and its measurement has come a long way in the last century, there are still many unknowns, particularly with regard to the causes of individual differences in cognitive ability measures. Understanding the factors influencing measurement of cognitive abilities is important in ensuring that the conclusions we make regarding test performance are valid, and in ensuring these tests are used appropriately. The present thesis has contributed to furthering understanding surrounding the measurement of Gf, the implications of the stimuli used to measure this ability, and sex differences found.

Researchers and practitioners who choose to use the RPM as a measure of inductive reasoning should be aware that the measure is somewhat contaminated by Gv, and that it also tends to show a male advantage, despite a lack of evidence for a male advantage on latent Gf. Additionally, they should be aware that the stimuli used to measure inductive reasoning may influence the construct measured, at least to the extent that it displays a sex differences in scores. This suggests that caution should be used in interpreting the scores of a single inductive reasoning measure, particularly as evidence of a sex difference in the ability measured.

# References

* Indicates paper included in meta-analysis; ** Indicates data obtained from author regarding sample and measures described in cited paper

*Abad, F. J., Colom, R., Rebollo, I., & Escorial, S. (2004). Sex differential item functioning in the Raven's Advanced Progressive Matrices: Evidence for bias. *Personality and Individual Differences, 36*, 1459-1470. doi:10.1016/S0191-8869(03)00241-1

**Abad, F. J., Sorrel, M. A., Roman, F. J., & Colom, R. (2016). The relationships between WAIS-IV factor index scores and educational level: A bifactor model approach. *Psychological Assessment, 28*, 987-1000. doi:10.1037/pas0000228

Abdel-Khalek, A. M. (1988). Egyptian results on the Standard Progressive Matrices. *Personality and Individual Differences, 9*, 193-195. doi:10.1016/0191-8869(88)90051-7

*Abdel-Khalek, A. M., & Lynn, R. (2009). Norms and sex differences for intelligence in Saudi Arabia assessed by the Standard Progressive Matrices. *Mankind Quarterly, 50*, 106-113. Retrieved from http://mankindquarterly.org/

*Abdel-Khalek, A. M., & Lynn, R. (2016). Sex difference in the intelligence of students at Alexandria University, Egypt. *Mankind Quarterly, 55*, 129-135. Retrieved from http://mankindquarterly.org/

*Abdel-Khalek, A. M., Nour-Eddin, A. S., & Lynn, R. (2014). A study of the intelligence of university students in Egypt. *Mankind Quarterly, 55*, 129-135. Retrieved from http://mankindquarterly.org/

*Abdel-Khalek, A. M., Nour-Eddin, A. S., & Lynn, R. (2015). A study of the performance of Egyptian college students on the Advanced Progressive Matrices. *Personality and Individual Differences*, *72*, 141-142. doi:10.1016/j.paid.2014.08.036

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2002). Individual differences in working memory within a nomological network of cognitive and perceptual speed abilities. *Journal of Experimental Psychology: General, 131*, 567-589. doi:10.1037//0096-3445.131.4.567

Adams, R. J., & Khoo, S. T. (1993). *Quest: The interactive test analysis system*. Melbourne, Australia: ACER.

Adams, R. J., Wu, M., & Wilson, M. (2012). *ACER ConQuest 3.1*. Melbourne, Australia: ACER.

*Ahmad, R., Khanum, S. J., Riaz, Z., & Lynn, R. (2008). Gender differences in means and variance on the Standard Progressive Matrices in Pakistan. *Mankind Quarterly, 49*, 50-57. Retrieved from http://mankindquarterly.org/

Al-Shahomee, A. A. (2012). A standardisation of the Standard Progressive Matrices for adults in Libya. *Personality and Individual Differences, 53*, 142-146. doi:10.1016/j.paid.2011.12.042

*Al-Shahomee, A. A., & Lynn, R. (2010). IQs of men and women and of arts and science students in Libya. *Mankind Quarterly, 51*, 153-157. Retrieved from http://mankindquarterly.org/

*Al-Shahomee, A. A., & Lynn, R. (2012). A standardisation of the Standard Progressive Matrices for Libyan adults aged 38 to 50 years. *Mankind Quarterly, 52*, 292-310. Retrieved from http://mankindquarterly.org/

Alderton, D. L., & Larson, G. E. (1990). Dimensionality of Raven's Advanced Progressive Matrices items. *Educational and Psychological Measurement, 50,* 887-900. doi:10.1177/0013164490504019

*Alexopoulos, D. S. (1996). Sex differences and I.Q. *Personality and Individual Differences*, 20, 445-450. doi:10.1016/0191-8869(95)00187-5

*Ali, M. S., Suliman, M. I., Kareem, A., & Iqbal, M. (2009). Comparison of gender performance on an intelligence test among medical students. *Journal of Ayub Medical College, 21*, 163-165. Retrieved from http://www.ayubmed.edu.pk/ JAMC/PAST/21-3?Sohail.pdf

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103*, 111-123. doi:10.1037/0033-2909.103.3.411

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

Arendasy, M. E., & Sommer, M. (2012). Gender differences in figural matrices: The moderating role of item design features. *Intelligence, 40,* 584-597. doi:10.1016/j.intell.2012.08.003

Arthur, W., & Day, D. V. (1994). Development of a short-form for the Raven Advanced Progressive Matrices test. *Educational and Psychological Measurement, 54,* 394-403. doi:10.1177/0013164494054002013

Arthur, W., Tubre, T. C., Paul, D. S., & Sanchez-Ku, M. L. (1999). College-sample psychometric and normative data on a short form of the Raven Advanced

Progressive Matrices test. *Journal of Psychoeducational Assessment, 17*, 354-361. doi:10.1177/073428299901700405

Arthur, 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. doi:10.1177/0013164493053002016

Asparouhov, T., & Muthén, B. (2006). *Robust chi square difference testing with mean and variance adjusted test statistics*. Retrieved from Mplus website: http://www.statmodel.com/download/webnotes/webnote10.pdf

Baddeley, A., Emslie, H., & Nimmo-Smith, I. (1993). The Spot-the-Word test: A robust estimate of verbal intelligence based on lexical decision. *British Journal of Clinical Psychology, 32*, 55-65. doi:10.1016/S0887-6177(99)00020-7

*Bakhiet, Al-Qudah, Essa, Y. A. S., Cheng, C. Y., & Lynn, R. (2016). Sex differences in the intelligence of university engineering students in Sudan. *Mankind Quarterly, 57*, 95-98. Retrieved from http://mankindquarterly.org/

*Bakhiet, Essa, Y. A. S., Abdelrasheed, N. S. G., Cheng, C. Y., & Lynn, R. (2016). Sex differences in the intelligence of university students in Thailand. *Mankind Quarterly, 57*, 72-74. Retrieved from http://mankindquarterly.org/

*Bakhiet, S. F. A., Haseeb, B. W. M., Seddieg, I. F., Cheng, H. L., & Lynn, R. (2015). Sex differences on Raven's Standard Progressive Matrices among 6 to 18 year olds in Sudan. *Intelligence, 50*, 10-13. doi:10.1016/j.intell.2015.01.013

*Batey, M., Furnham, A., & Safiullina, X. (2010). Intelligence, general knowledge and personality as predictors of creativity. *Learning and Individual Differences, 20*, 532-535. doi:10.1016/j.lindif.2010.04.008

Bennett, G. K., Seashore, H. G., & Wesman, A. G. (1989). *Australian manual for the Differential Aptitude Tests: Forms V and W.* Australia: The Psychological Corporation.

*Birkett, P. (1980). Predicting spatial ability from hemispheric 'non-verbal' lateralisation: Sex, handedness and task differences implicate encoding strategy effects. *Acta Psychologica, 46,* 1-14. doi:10.1016/0001-6918(80)90056-6

Blinkhorn, S. (2005). Intelligence: a gender bender. *Nature, 438,* 31-32. doi:10.1038/438031a

Blinkhorn, S. S. (2006). Intelligence: Is there a sex difference in IQ scores? (Reply). *Nature, 442,* E1-E2. doi:10.1038/nature04967

Bors, D. A., & Forrin, B. (1995). Age, speed of information-processing, recall, and fluid intelligence. *Intelligence, 20,* 229-248. doi:10.1016/0160-2896(95)90009-8

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,* 382-398. doi:10.1177/0013164498058003002

Bors, D. A., & Vigneau, F. (2001). The effect of practice on Raven's Advanced Progressive Matrices. *Learning and Individual Differences, 13,* 291-312. doi:10.1016/S1041-6080(03)00015-3

Borst, G., & Kosslyn, S. M. (2010). Individual differences in spatial mental imagery. *Quarterly Journal of Experimental Psychology, 63,* 2031-2050. doi:10.1080/17470211003802459

Botella, J., Pena, D., Contreras, M. J., Shih, P. C., & Santacreu, J. (2009). Performance as a function of ability, resources invested, and strategy used. *Journal of General Psychology, 136,* 41-69. doi:10.3200/GENP.136.1.41-70

Boulter, D. R., & Kirby, J. R. (1994). Identification of strategies used in solving transformational geometry problems. *Journal of Educational Research, 87,* 298-303. doi:10.1080/00220671.1994.9941257

*Britton, A., Singh-Manoux, A., & Marmot, M. (2004). Alcohol consumption and cognitive function in the Whitehall II Study. *American Journal of Epidemiology, 160,* 240-247. doi:10.1093/aje/kwh206

*Bromley, D. B. (1991). Aspects of written language production over adult life. *Psychology and Aging, 6,* 296-308. doi:10.1037/0882-7974.6.2.296

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park CA: Sage.

Bucik, V., & Neubauer, A. C. (1996). Bimodality in the Berlin model of intelligence structure (BIS): A replication study. *Personality and Individual Differences, 21,* 987-1005. doi:10.1016/S0191-8869(96)00129-8

Bui, M., & Birney, D. P. (2014). Learning and individual differences in Gf processes and Raven's. *Learning and Individual Differences, 32,* 104-113. doi:10.1016/j.lindif.2014.03.008

Burgaleta, M., Head, K., Álvarez-Linera, J., Martínez, K., Escorial, S., Haier, R., & Colom, R. (2012). Sex differences in brain volume are related to specific skills, not to general intelligence. *Intelligence, 40,* 60-68. doi:10.1016/j.intell.2011.10.006

Burke, H. R. (1958). Raven's Progressive Matrices: A review and critical evaluation. *Journal of Genetic Psychology, 93,* 199-228. doi:10.1080/00221325.1958.10532420

Burns, N. R., Bastian, V. A., & Nettelbeck, T. (2007). Emotional intelligence: More than personality and cognitive ability? In G. Matthews, M. Zeidner & R. D. Roberts (Eds.), *The science of emotional intelligence: Knowns and unknowns* (pp. 167-196). New York, NY: Oxford University Press.

Burns, K. M., Burns, N. R., & Ward, L. (2016). Confidence-More a personality or ability trait? It depends on how it is measured: A comparison of young and older adults. *Frontiers in Psychology, 7*, 518. doi:10.3389/fpsyg.2016.00518

Burns, N. R., Nettelbeck, T., McPherson, J., & Stankov, L. (2007). Perceptual learning on inspection time and motion perception. *Journal of General Psychology, 134,* 83-100. doi:10.3200/GENP.134.1.83-100

Calvin, C. M., Fernandes, C., Smith, P., Visscher, P. M., & Deary, I. J. (2010). Sex, intelligence and educational achievement in a national cohort of over 175,000 11-year-old schoolchildren in England. *Intelligence, 38*, 424-432. doi:10.1016/j.intell.2010.04.005

Camarata, S., & Woodcock, R. (2006). Sex differences in processing speed: Developmental effects in males and females. *Intelligence, 34*, 231-252. doi:10.1016/j.intell.2005.12.001

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*, 404-431. doi:10.1037/0033-295x.97.3.404

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Victoria, Australia: Cambrige University Press.

Carroll, J. B. (2003). The higher-stratum structure of cognitive abilities: Current evidence supports g and about ten broad factors. In H. Nyborg (Ed.), *The scientific study of general intelligence: Tribute to Arthur R. Jensen* (pp. 5-21). The Netherlands: Pergamon Press.

Cattell, R. B. (1941). Some theoretical issues in adult intelligence testing. *Psychology Bulletin, 38*, 592.

Cattell, R. B. (1971). *Abilities: Their structure, growth, and action*. New York, NY: North Holland.

Cattell, R. B. (1980). They talk of some strict testing of us-Pish. *Behavioral and Brain Sciences, 3*, 336-337. doi:10.1017/S0140525X00005203

Cattell, R. B. (1987). *Intelligence: Its structure, growth, and action*. New York, NY: North Holland.

*Chamorro-Premuzic, T., & Arteche, A. (2008). Intellectual competence and academic performance: Preliminary validation of a model. *Intelligence, 36*, 564-573. doi:10.1016/j.intell.2008.01.001

*Chamorro-Premuzic, T., Moutafi, J., & Furnham, A. (2005). The relationship between personality traits, subjectively-assessed and fluid intelligence. *Personality and Individual Differences, 38*, 1517-1528. doi:10.1016/j.paid.2004.09.018

*Chamorro-Premuzic, T., Quiroga, M. A., & Colom, R. (2009). Intellectual competence and academic performance: A Spanish study. *Learning and Individual Differences, 19*, 486-491. doi:10.1016/j.lindif.2009.05.002

Chiesi, F., Ciancaleoni, M., Galli, S., Morsanyi, K., & Primi, C. (2012). Item response theory analysis and differential item functioning across age, gender and country of a short form of the Advanced Progressive Matrices. *Learning and Individual Differences, 22*, 390-396. doi:10.1016/j.lindif.2011.12.007

Chiesi, F., Ciancaleoni, M., Galli, S., & Primi, C. (2012). Using the Advanced Progressive Matrices (Set I) to assess fluid ability in a short time frame: An item response theory-based analysis. *Psychological Assessment, 24*, 892-900. doi:10.1037/a0027830

*Choi, J., & L'Hirondelle, N. (2005). Object location memory: A direct test of the verbal memory hypothesis. *Learning and Individual Differences, 15*, 237-245. doi:10.1016/j.lindif.2005.02.001

Chuderski, A. (2015). The broad factor of working memory is virtually isomorphic to fluid intelligence tested under time pressure. *Personality and Individual Differences, 85*, 98-104. doi:10.1016/j.paid.2015.04.046

Cikrikci-Demirtasli, N. (2000). *A study of Raven Standard Progressive Matrices test's item measures under classic and item response models: An empirical comparison*. Paper presented at the 31st European Mathematical Psychology Congress, Graz.

Cockcroft, K., & Israel, N. (2011). The Raven's Advanced Progressice Matrices: A comparison of relationships with verbal ability tests. *South African Journal of Psychology, 41*, 363-372. doi:10.1177/008124631104100310

Codorniu-Raga, M. J., & Vigil-Colet, A. (2003). Sex differences in psychometric and chronometric measures of intelligence among young adolescents. *Personality and Individual Differences, 35*, 681-689. doi:10.1016/S0191-8869(02)00245-3

Colom, R., & Abad, F. J. (2007). Advanced Progressive Matrices and sex differences: Comment to Mackintosh and Bennett (2005). *Intelligence, 35*, 183-185. doi:10.1016/j.intell.2006.06.003

**Colom, R., Abad, F. J., Quiroga, M. Á., Shih, P. C., & Flores-Mendoza, C. (2008). Working memory and intelligence are highly related constructs, but why? *Intelligence, 36*, 584-606. doi:10.1016/j.intell.2008.01.002

Colom, R., Contreras, J., Arend, I., Leal, O. G., & Santacreu, J. (2004). Sex differences in verbal reasoning are mediated by sex differences in spatial ability.

*Psychological Record, 54,* 365-372. Retrieved from http://opensiuc.lib.siu.edu/tpr/

**Colom, R., Contreras, M. J., Botella, J., & Santacreu, J. (2002). Vehicles of spatial ability. *Personality and Individual Differences, 32,* 903-912. doi:10.1016/S0191-8869(01)00095-2

*Colom, R., Escorial, S., & Rebollo, I. (2004). Sex differences on the Progressive Matrices are influenced by sex differences on spatial ability. *Personality and Individual Differences, 37*, 1289-1293. doi:10.1016/j.paid.2003.12.014

*Colom, R., & Garcia-Lopez, O. (2002). Sex differences in fluid intelligence among high school graduates. *Personality and Individual Differences, 32*, 445-451. doi:10.1016/S0191-8869(01)00040-X

*Colom, R., Juan-Espinosa, M., Abad, F., & Garcia, L. F. (2000). Negligible sex differences in general intelligence. *Intelligence*, *28*, 57-68. doi:10.1016/S0160-2896(99)00035-5

*Colom, R., & Lynn, R. (2004). Testing the developmental theory of sex differences in intelligence on 12-18 year olds. *Personality and Individual Differences, 36,* 75-82. doi:10.1016/S0191-8869(03)00053-9

**Colom, R., Privado, J., García, L. F., Estrada, E., Cuevas, L., & Shih, P.-C. (2015). Fluid intelligence and working memory capacity: Is the time for working on intelligence problems relevant for explaining their large relationship? *Personality and Individual Differences, 79*, 75-80. doi:10.1016/j.paid.2015.01.051

**Colom, R., & Quiroga, M. A. (2009). Neuroticism, intelligence, and intra-individual variability in elementary cognitive tasks: Testing the mental noise hypothesis. *Psicothema, 21*, 403-408. Retrieved from http://www.psicothema.com

*Colom, R., Quiroga, M. A., & Juan-Espinosa, M. (1999). Are cognitive sex differences disappearing? Evidence from Spanish populations. *Personality and Individual Differences, 27,* 1189-1195. doi:10.1016/S0191-8869(99)00062-8

Colom, R., Román, F. J., Abad, F. J., Shih, P. C., Privado, J., Froufe, M., . . . Jaeggi, S. M. (2013). Adaptive n-back training does not improve fluid intelligence at the construct level: Gains on individual tests suggest that training may enhance visuospatial processing. *Intelligence, 41*, 712-727. doi:10.1016/j.intell.2013.09.002

*Colom, R., Stein, J. L., Rajagopalan, P., Martinez, K., Hermel, D., Wang, Y., . . . Thompson, P. M. (2013). Hippocampal structure and human cognition: Key role of spatial processing and evidence supporting the efficiency hypothesis in females. *Intelligence, 41*, 129-140. doi:10.1016/j.intell.2013.01.002

Court, J. H. (1983). Sex differences in performance on Raven's Progressive Matrices: A review. *The Alberta Journal of Educational Research, 16*, 54-74.

Crawford, J. D. (1991). The relationship between tests of sustained attention and fluid intelligence. *Personality and Individual Differences, 12*, 599-611. doi:10.1016/0191-8869(91)90257-C

*Čvorović, J., & Lynn, R. (2014). Sex differences in intelligence: Some new data from Serbia. *Mankind Quarterly, 55*(1-2), 101-109. Retrieved from http://mankindquarterly.org/

Đapo, N., & Kolenović-Đapo, J. (2012). Sex differences in fluid intelligence: Some findings from Bosnia and Herzegovina. *Personality and Individual Differences, 53,* 811-815. doi:10.1016/j.paid.2012.05.036

*Deary, I. J., Der, G., & Ford, G. (2001). Reaction times and intelligence differences: A population-based cohort study. *Intelligence, 29,* 389-399. doi:10.1016/S0160-2896(01)00062-9

Deary, I. J., Irwing, P., Der, G., & Bates, T. C. (2007). Brother–sister differences in the g factor in intelligence: Analysis of full, opposite-sex siblings from the NLSY1979. *Intelligence, 35,* 451-456. doi:10.1016/j.intell.2006.09.003

Del Re, A. C. (2013). compute.es: Compute Effect Sizes (R Package version 0.2-4). Retrieved from: http://cran.r-project.org/web/packages/compute.es

*Der, G., Batty, G. D., Benzeval, M., Deary, I. J., Green, M. J., McGlynn, L., . . . Shiels, P. G. (2012). Is telomere length a biomarker for aging? Cross-sectional evidence from the west of Scotland. *PLoS ONE, 7,* e45166. doi:10.1371/journal.pone.0045166

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. doi:10.1016/0160-2896(95)90023-3

*Diaz, & Lynn, R. (2016). Sex differences on the WAIS-IV in Chile. *Mankind Quarterly, 57,* 52-57. Retrieved from http://mankindquarterly.org/

*Díaz, A., Sellami, K., Infanzón, E., Lanzón, T., & Lynn, R. (2010). Sex differences in means and variance of intelligence: Some data for Spain. *Mankind Quarterly, 50,* 210-220. Retrieved from http://mankindquarterly.org/

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. doi:10.1177/001316448104100438

*Dolan, C. V., Colom, R., Abad, F. J., Wicherts, J. M., Hessen, D. J., & van de Sluis, S. (2006). Multi-group covariance and mean structure modeling of the relationship between the WAIS-III common factors and sex educational attainment in Spain. *Intelligence, 34,* 193-210. doi:10.1016/j.intell.2005.09.003

Dykiert, D., Gale, C. R., & Deary, I. J. (2009). Are apparent sex differences in mean IQ scores created in part by sample restriction and increased male variance? *Intelligence, 37,* 42-47. doi:10.1016/j.intell.2008.06.002

Ebisch, S. J., Perrucci, M. G., Mercuri, P., Romanelli, R., Mantini, D., Romani, G. L., . . . Saggino, A. (2012). Common and unique neuro-functional basis of induction, visualization, and spatial relationships as cognitive components of fluid intelligence. *NeuroImage, 62,* 331-342. doi:10.1016/j.neuroimage.2012.04.053

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315*, 629-634. doi:10.1136/bmj.315.7109.629

Eid, M., Lischetzke, T., Nussbeck, F. W., & Trierweiler, L. I. (2003). Separating trait effects from trait-specific method effects in multitrait-multimethod models: A multiple-indicator CT-C(M-1) model. *Psychological Methods, 8,* 38-60. doi:10.1037/1082-989X.8.1.38

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

*Escorial, S., Juan-Espinosa, M., Garcia, L., Rebollo, I., & Colom, R. (2003). Does g variance change in adulthood? Testing the age de-differentiation hypothesis across sex. *Personality and Individual Differences, 34,* 1525-1532. doi:10.1016/s0191-8869(02)00133-2

*Essa, Y. A. S., Abdelrasheed, N. S. G., Bakhiet, S. F. A., Cheng, H., Dwieb, A. M. M., & Lynn, R. (2016). Sex differences in the intelligence of students at an Egyptian university. *Personality and Individual Differences, 95,* 183-184. doi:10.1016/j.paid.2016.02.036

**Estrada, E., Ferrer, E., Abad, F. J., Roman, F. J., & Colom, R. (2015). A general factor of intelligence fails to account for changes in tests' scores after cognitive practice: A longitudinal multi-group latent-variable study. *Intelligence, 50,* 93-99. doi:10.1016/j.intell.2015.02.004

Feingold, A. (1988). Cognitive gender differences are disappearing. *American Psychologist, 43*, 95-103. doi:10.1037/0003-066x.43.2.95

*Flores-Mendoza, C., Darley, M., & Fernandes, H. B. F. (2016). Cognitive sex differences in Brazil. *Mankind Quarterly, 57*, 34-51. Retrieved from http://mankindquarterly.org/

Flores-Mendoza, C., Widaman, K. F., Rindermann, H., Primi, R., Mansur-Alves, M., & Pena, C. C. (2013). Cognitive sex differences in reasoning tasks: Evidence from Brazilian samples of educational settings. *Intelligence, 41*, 70-84. doi:10.1016/j.intell.2012.11.002

*Flynn, J. R., & Rossi-Case, L. (2011). Modern women match men on Raven's Progressive Matrices. *Personality and Individual Differences, 50,* 799-803. doi:10.1016/j.paid.2010.12.035

*Foley, W. J., & Proff, F. C. (1965). NDEA Institute trainees and vocational-rehabilitation counselors: A comparison of characteristics. *Counselor Education and Supervision, 4*, 154-159. doi:10.1002/j.1556-6978.1965.tb02164.x

Fugard, A. J. B., Stewart, M. E., & Stenning, K. (2011). Visual/verbal-analytic reasoning bias as a function of self-reported autistic-like traits: A study of typically developing individuals solving Raven's Advanced Progressive Matrices. *Autism, 15,* 327-340. doi:10.1177/1362361310371798

*Furnham, A., Taylor, J., & Chamorro-Premuzic, T. (2008). Personality and intelligence correlates of assessment center exercises. *Individual Differences Research, 6*, 181-192.

*Gale, C. R., Sayer, A. A., Cooper, C., Dennison, E. M., Starr, J. M., Whalley, L. J., . . . Team, H. A. S. (2011). Factors associated with symptoms of anxiety and depression in five cohorts of community-based older people: the HALCyon (Healthy Ageing across the Life Course) Programme. *Psychological Medicine, 41*, 2057-2073. doi:10.1017/S0033291711000195

*Gangestad, S. W., Thornhill, R., & Garver-Apgar, C. E. (2010). Men's facial masculinity predicts changes in their female partners' sexual interests across the ovulatory cycle, whereas men's intelligence does not. *Evolution and Human Behavior, 31*, 412-424. doi:10.1016/j.evolhumbehav.2010.06.001

Geary, D. C., Saults, S. J., Liu, F., & Hoard, M. K. (2000). Sex differences in spatial cognition, computational fluency, and arithmetical reasoning. *Journal of Experimental Child Psychology, 77*, 337-353. doi:10.1006/jecp.2000.2594

Geiser, C., Lehmann, W., & Eid, M. (2006). Separating "Rotators" from "Nonrotators" in the Mental Rotations Test: A multigroup latent class analysis. *Multivariate Behavioral Research, 41*, 261-293. doi:10.1207/s15327906mbr4103_2

Genesys Australia. (2012). GeneSys Australian Norms 2012: Author.

Genesys Australia. (2015). [General Reasoning Test 2]. Unpublished raw data.

Gibson, W. A. (1960). Nonlinear factors in two dimensions. *Psychometrika, 25*, 381-392. doi:10.1007/BF02289755

Gignac, G. E. (2007). Working memory and fluid intelligence are both identical to g?! Reanalyses and critical evaluation. *Psychology Science, 49*, 187-207.

Gignac, G. E. (2015). Raven's is not a pure measure of general intelligence: Implications for g factor theory and the brief measurement of g. *Intelligence, 52*, 71-79. doi:10.1016/j.intell.2015.07.006

Gonthier, C., & Thomassin, N. (2015). Strategy use fully mediates the relationship between working memory capacity and performance on Raven's Matrices. *Journal of Experimental Psychology, 144*, 916-924. doi:10.1037/xge0000101

Grigoriev, A., & Lynn, R. (2014). A study of the intelligence of Kazakhs, Russians and Uzbeks in Kazakhstan. *Intelligence, 46*, 40-46. doi:10.1016/j.intell.2014.05.004

Gromping, U. (2006). Relative importance for linear regression in R: The package relaimpo. *Journal of Statistical Software, 17*, 1-27. Retrieved from https://www.jstatsoft.org

Guay, R. B., McDaniel, E., & Angelo, S. (1978). *Analytic factor confounding spatial ability measurement*. Paper presented at the American Psychological Association, Toronto, Canada.

Gustafsson, J. E. (1984). A unifying model for the structure of intellectual abilities. *Intelligence, 8*, 179-203. doi:10.1016/0160-2896(84)90008-4

Habeck, C., Steffener, J., Barulli, D., Gazes, Y., Razlighi, Q., Shaked, D., . . . Stern, Y. (2015). Making cognitive latent variables manifest: Distinct neural networks for fluid reasoning and processing speed. *Journal of Cognitive Neuroscience, 27*, 1249-1258. doi:10.1162/jocn_a_00778

*Hakstian, A. R., & Cattell, R. B. (1982). *Manual for the Comprehensive Ability Battery*. Champaign, Illinois: Institute for Personality and Ability Testing.

Hakstian, A. R., & Cattell, R. B. (1975a). *The Comprehensive Ability Battery*. Champaign, IL: Institute for Personality and Ability Testing.

Hakstian, A. R., & Cattell, R. B. (1975b). An examination of adolescent sex differences in some ability and personality traits. *Canadian Journal of Behavioural Science, 7*, 295-312.

Halpern, D. F. (2012). *Sex differences in cognitive abilities*. New York: NY: Psychology Press.

Hambrick, D. Z., & Altmann, E. M. (2015). The role of placekeeping ability in fluid intelligence. *Psychonomic Bulletin and Review, 22*, 1104-1110. doi:10.3758/s13423-014-0764-5

*Hambrick, D. Z., Oswald, F. L., Darowski, E. S., Rench, T. A., & Brou, R. (2010). Predictors of multitasking performance in a synthetic work paradigm. *Applied Cognitive Psychology, 24*, 1149-1167. doi:10.1002/acp.1624

*Hambrick, D. Z., Pink, J. E., Meinz, E. J., Pettibone, J. C., & Oswald, F. L. (2008). The roles of ability, personality, and interests in acquiring current events knowledge: A longitudinal study. *Intelligence, 36*, 261-278. doi:10.1016/j.intell.2007.06.004

Harrison, T. L., Shipstead, Z., & Engle, R. W. (2015). Why is working memory capacity related to matrix reasoning tasks? *Memory and Cognition, 43*, 389-396. doi:10.3758/s13421-014-0473-3

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*, 139-164. doi:10.1177/014662168500900204

*Hattori, K., & Lynn, R. (1997). Male-female differences on the Japanese WAIS-R. *Personality and Individual Differences, 23*, 531-533. doi:10.1016/S0191-8869%2897%2900055-X

*Hegarty, M., Keehner, M., Khooshabeh, P., & Montello, D. R. (2009). How spatial abilities enhance, and are enhanced by, dental education. *Learning and Individual Differences, 19*, 61-70. doi:10.1016/j.lindif.2008.04.006

Heil, M., & Jansen-Osmann, P. (2008). Sex differences in mental rotation with polygons of different complexity: Do men utilize holistic processes whereas women prefer piecemeal ones? *Quarterly Journal of Experimental Psychology, 61*, 683-689. doi:10.1080/17470210701822967

Hertzog, C., & Carter, L. (1982). Sex differences in the structure of intelligence: A confirmatory factor-analysis. *Intelligence, 6*, 287-303. doi:10.1016/0160-2896(82)90005-8

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistical Methods, 21*, 1539-1558. doi:10.1002/sim.1186

Horn, J. L. (1965). *Fluid and crystallized intelligence: A factor analytic and developmental study of the structure among primary mental abilities* (Unpublished doctoral dissertation). University of Illinois, Champaign.

Horn, J. L. (1988). Thinking about human abilities. In J. R. Nesselroade (Ed.), *Handbook of multivariate psychology* (pp. 645-685). New York, NY: Academic Press.

Horn, J. L. (1994). The theory of fluid and crystallized intelligence. In R. J. Sternberg (Ed.), *The encyclopedia of intelligence* (pp. 443-451). New York, NY: Macmillan.

Horn, J. L., & Blankson, N. (2005). Foundations for better understanding of cognitive abilities. In D. P. Flanagan & K. S. McGrew (Eds.), *Contemporary intellectual assessment* (pp. 41-68). New York, NY: Guilford Press.

Horn, W. (1983). *Leistungsprufsystem [Performance-Test-System]*. (2nd ed.). Gottingen: Hogrefe.

House, J. D., & Keeley, E. J. (1995). Gender bias in prediction of graduate grade performance from Miller Analogies Test scores. *Journal of Psychology, 129*, 353-355. doi:10.1080/00223980.1995.9914972

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling-a Multidisciplinary Journal, 6*, 1-55. doi:10.1080/10705519909540118

Hungi, N. (2005). Applying the Rasch model to detect biased items. In S. Alagumalai, D. D. Curtis, & N. Hungi (Eds.), *Applied Rasch measurement* (pp. 139-158). The Netherlands: Springer.

Hunt, E. (1974). Quote the raven? Nevermore! In G. W. Gregg (Ed.), *Knowledge and cognition*. Hillsdale, NJ: Erlbaum.

Hunt, E. (1999). Intelligence and human resources: Past, present and future. In P. L. Ackerman, P. C. Killeen, & R. D. Roberts (Eds.), *Learning and individual differences: Process, trait and content determinants* (pp. 3-28). Washington, DC: American Psychological Association.

Hunt, E. (2011). *Human intelligence*. New York: Cambridge University Press.

Hunt, E., & Madhyastha, T. (2008). Recruitment modeling: An analysis and an application to the study of male-female differences in intelligence. *Intelligence, 36*, 653-663. doi:10.1016/j.intell.2008.03.002

Hunter, J. E. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. *Journal of Vocational Behavior, 29*, 340-362. doi:10.1016/0001-8791(86)90013-8

Hyde, J. S., & Linn, M. C. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin, 104*, 53-69. doi:10.1037/0033-2909.104.1.53

Irwing, P. (2015). Sex differences in g: An analysis of the US standardization sample of the WAIS-III. *Personality and Individual Differences, 53*, 126-131. doi:10.1016/j.paid.2011.05.001

Irwing, P., & Lynn, R. (2005). Sex differences in means and variability on the Progressive Matrices in university students: A meta-analysis. *British Journal of Psychology, 96,* 505-524. doi:10.1348/000712605X53542

Irwing, P., & Lynn, R. (2006). Intelligence: Is there a sex difference in IQ scores? *Nature, 442*, E1. doi:10.1038/nature04966

Jackson, S. A., Kleitman, S., Stankov, L., & Howie, P. (2016). Individual differences in decision making depend on cognition, monitoring and control. *Journal of Behavioral Decision Making*. Advance online publication. doi:10.1002/bdm.1939

Jäger, A. O. (1982). Mehrmodale klassifikation von inteligenzleistungen: Experimentell kontrollierte weiterentwicklung eines deskripriven intelligenzstruckturmodells. *Diagnostica, 28,* 195-225.

Jaušovec, N., & Jaušovec, K. (2012). Sex differences in mental rotation and cortical activation patterns: Can training change them? *Intelligence, 40*, 151-162. doi:10.1016/j.intell.2012.01.005

Jensen, A. R. (1998). *The g factor: The science of mental ability.* Westport, CT: Praeger.

Johnson, W., & Bouchard, T. J. (2005a). Constructive replication of the visual-perceptual-image rotation model in Thurstone's (1941) battery of 60 tests of mental ability. *Intelligence, 33*, 417-430. doi:10.1016/j.intell.2004.12.001

Johnson, W., & Bouchard, T. (2005b). The structure of human intelligence: It is verbal, perceptual, and image rotation (VPR), not fluid and crystallized. *Intelligence, 33*, 393-416. doi:10.1016/j.intell.2004.12.002

**Johnson, W., & Bouchard, T. J. (2007). Sex differences in mental abilities: g masks the dimensions on which they lie. *Intelligence, 35,* 23-39. doi:10.1016/j.intell.2006.03.012

Johnson, W., Carothers, A., & Deary, I. J. (2008). Sex differences in variability in general intelligence: A new look at the old question. *Perspectives on Psychological Science, 3,* 518-531. doi:10.1111/j.1745-6924.2008.00096.x

*Kagan, D. M., & Stock, W. A. (1980). Equivalencing MAT and GRE scores using simple linear transformation and regression methods. *Journal of Experimental Education, 49,* 34-37. doi:10.1080/00220973.1980.11011759

*Kaufman, A. S., & Horn, J. L. (1996). Age changes on tests of fluid and crystallized ability for women and men on the Kaufman Adolescent and Adult Intelligence Test (KAIT) at ages 17-94 years. *Archives of Clinical Neuropsychology, 11,* 97-121. doi:10.1016/0887-6177(95)00003-8

*Kaufman, A. S., & Kaufman, N. L. (2004). *Kaufman Brief Intelligence Test, Second Edition*. Bloomington, MN: Pearson, Inc.

**Keith, T. Z., Reynolds, M. R., Patel, P. G., & Ridley, K. P. (2008). Sex differences in latent cognitive abilities ages 6 to 59: Evidence from the Woodcock-Johnson III tests of cognitive abilities. *Intelligence, 36,* 502-525. doi:10.1016/j.intell.2007.11.001

Keith, T. Z., Reynolds, M. R., Roberts, L. G., Winter, A. L., & Austin, C. A. (2011). Sex differences in latent cognitive abilities ages 5 to 17: Evidence from the

Differential Ability Scales-Second Edition. *Intelligence, 39*, 389-404. doi:10.1016/j.intell.2011.06.008

Kelly, S. J., Burns, N. R., Bradman, G., Wittert, G., and Daniel, M. (2012). Does IQ vary systematically with all measures of socioeconomic status in a cohort of middle-aged, and older, men? *Sociology Mind, 2*, 394-400. doi:10.4236/sm.2012.24052

Kenny, D. A. (1976). An empirical application of confirmatory factor analysis to the multitrait-multimethod matrix. *Journal of Experimental Social Psychology, 12*, 247-252. doi:10.1016/0022-1031(76)90055-X

Kettner, G., Seashore, B. H. G., & Wesman, A. G. (1966). *Differential aptitude tests: manual. Forms L and M* (4th ed.). New York, NY: Psychological Corporation.

*Khaleefa, O., Ali, K., & Lynn, R. (2010). IQ and head size in a sample in Sudan. *Mankind Quarterly, 51*, 108-111. Retrieved from http://mankindquarterly.org/

*Khaleefa, O., Amer, Z., & Lynn, R. (2014). IQ differences between arts and science students at the university of Khartoum. *Mankind Quarterly, 55*, 136-146. Retrieved from http://mankindquarterly.org/

*Khaleefa, O., Khatib, M. A., Mutwakkil, M. M., & Lynn, R. (2008). Norms and gender differences on the Progressive Matrices in Sudan. *Mankind Quarterly, 49*, 176-182. Retrieved from http://mankindquarterly.org/

*Khaleefa, O., & Lynn, R. (2008). Sex differences on the progressive matrices: Some data from Syria. *Mankind Quarterly, 48*, 345-351. Retrieved from http://mankindquarterly.org/

Kirby, J. R., & Lawson, M. J. (1983). Effects of strategy training on Progressive Matrices performance. *Contemporary Educational Psychology, 8*, 127-140. doi:10.1016/0361-476x(83)90004-8

Klein, R. M., Dilchert, S., Ones, D. S., & Dages, K. D. (2015). Cognitive predictors and age-based adverse impact among business executives. *Journal of Applied Psychology, 100*, 1497-1510. doi:10.1037/a0038991

Kleitman, S., & Stankov, L. (2007). Self-confidence and metacognitive processes. *Learning and Individual Differences*, *17*, 161-173. doi:10.1016/j.lindif.2007.03.004

Kline, P. (1994). *An easy guide to factor analysis*. New York, NY: Routledge.

Kline, R. B. (1998). *Principles and practice of structural equation modelling*. New York, NY: Guilford Press.

Kubinger, K. D., Formann, A. K., & Farkas, M. G. (1991). Psychometric shortcomings of Raven's Standard Progressive matrices, in particular for computerized testing. *Revue Européenne de Psychologie Appliquee, 41*, 295-300.

Kunda, M., McGreggor, K., & Goel, A. K. (2013). A computational model for solving problems from the Raven's Progressive Matrices intelligence test using iconic visual representations. *Cognitive Systems Research, 22-23,* 47-66. doi:10.1016/j.cogsys.2012.08.001

Kvist, A. V., & Gustafsson, J. E. (2008). The relation between fluid intelligence and the general factor as a function of cultural background: A test of Cattell's Investment Theory. *Intelligence, 36,* 422-436. doi:10.1016/j.intell.2007.08.004

Kyllonen, P. C., Lohman, D. F., & Snow, R. E. (1984). Effects of aptitudes, strategy training, and task facets on spatial task-performance. *Journal of Educational Psychology, 76*, 130-145. doi:10.1037//0022-0663.76.1.130

Lakin, J., & Gambrell, J. (2014). Sex differences in fluid reasoning: Manifest and latent estimates from the Cognitive Abilities Test. *Journal of Intelligence*, *2*, 36-55. doi:10.3390/jintelligence2020036

Lance, C. E., Noble, C. L., & Scullen, S. E. (2002). A critique of the correlated trait-correlated method and correlated uniqueness models for multitrait-multimethod data. *Psychological Methods, 7,* 228-244. doi:0.1037//1082-989X.7.2.228

Lang, J. W. B., Kersting, M., & Beauducel, A. (2016). Hierarchies of factor solutions in the intelligence domain: Applying methodology from personality psychology to gain insights into the nature of intelligence. *Learning and Individual Differences, 47,* 37-50. doi:10.1016/j.lindif.2015.12.003

Lemos, G. C., Abad, F. J., Almeida, L. S., & Colom, R. (2013). Sex differences on g and non-g intellectual performance reveal potential sources of STEM discrepancies. *Intelligence, 41,* 11-18. doi:10.1016/j.intell.2012.10.009

Liang, P., Jia, X., Taatgen, N. A., Zhong, N., & Li, K. (2014). Different strategies in solving series completion inductive reasoning problems: An fMRI and computational study. *International Journal of Psychophysiology, 93,* 253-260. doi:10.1016/j.ijpsycho.2014.05.006

Lim, T. K. (1994). Gender-related differences in intelligence: Application of confirmatory factor-analysis. *Intelligence, 19,* 179-192. doi:10.1016/0160-2896(94)90012-4

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16,* 878. Retrieved from www.rasch.org/rmt/rmt162f.htm

Lindberg, S. M., Hyde, J. S., Petersen, J. L., & Linn, M. C. (2010). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin, 136*, 1123-1135. doi:10.1037/a0021276

Linn, M. C., & Petersen, A. C. (1985). Emergence and characterization of sex differences in spatial ability: A meta-analysis. *Child Development, 56*, 1479-1498. doi:10.2307/1130467

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling, 9*, 151-173. doi:10.1207/S15328007SEM0902_1

Loesche, P., Wiley, J., & Hasselhorn, M. (2015). How knowing the rules affects solving the Raven Advanced Progressive Matrices test. *Intelligence, 48*, 58-75. doi:10.1016/j.intell.2014.10.004

Lohman, D. F., & Lakin, J. M. (2009). Consistencies in sex differences on the Cognitive Abilities Test across countries, grades, test forms, and cohorts. *British Journal of Educational Psychology, 79,* 389-407. doi:10.1348/000709908X354609

*Lynn, R. (1998). Sex differences in intelligence: Data from a Scottish standardisation of the WAIS-R. *Personality and Individual Differences, 24,* 289-290. doi:10.1016/S0191-8869(97)00165-7

Lynn, R. (1999). Sex differences in intelligence and brain size: A developmental theory. *Intelligence, 27,* 1-12. doi:10.1016/S0160-2896(99)00009-4

*Lynn, R. (2014). A study of intelligence in Cambodia. *Mankind Quarterly, 54*, 458-464. Retrieved from http://mankindquarterly.org/

Lynn, R., Allik, J., & Irwing, P. (2004). Sex differences on three factors identified in Raven's Standard Progressive Matrices. *Intelligence, 32*, 411-424. doi:10.1016/j.intell.2004.06.007

Lynn, R., Backhoff, E., & Contreras, L. A. (2005). Ethnic and racial differences on the Standard Progressive Matrices in Mexico. *Journal of Biosocial Science, 37*, 107-113. doi:10.1017/S0021932003006497

*Lynn, R., & Dai, X. Y. (1993). Sex differences on the Chinese standardization sample of the WAIS-R. *Journal of Genetic Psychology, 154*, 459-463. doi:10.1080/00221325.1993.9914744

*Lynn, R., & Hur, Y. M. (2016). Sex differences in the WAIS-IV on the South Korean standardization sample. *Mankind Quarterly, 57,* 58-65. Retrieved from http://mankindquarterly.org/

*Lynn, R., & Irwing, P. (2004a). Sex differences on the Advanced Progressive Matrices in college students. *Personality and Individual Differences, 37*, 219-223. doi:10.1016/j.paid.2003.08.028

Lynn, R., & Irwing, P. (2004b). Sex differences on the Progressive Matrices: A meta-analysis. *Intelligence, 32,* 481-498. doi:10.1016/j.intell.2040.06.008

*Lynn, R., & Tse-Chan, P. W. (2003). Sex differences on the Progressive Matrices: Some data from Hong Kong. *Journal of Biosocial Science*, *35*, 145-150. doi:10.1017/S0021932003001457

*MacCann, C. (2010). Further examination of emotional intelligence as a standard intelligence: A latent variable analysis of fluid intelligence, crystallized intelligence, and emotional intelligence. *Personality and Individual Differences, 49,* 490-496. doi:10.1016/j.paid.2010.05.010

Mackintosh, N. J. (1996). Sex differences and IQ. *Journal of Biosocial Science, 28,* 559-571. doi:10.1017/s0021932000022586

Mackintosh, N. J., & Bennett, E. S. (2005). What do Raven's Matrices measure? An analysis in terms of sex differences. *Intelligence, 33*, 663-674. doi:10.1016/j.intell.2005.03.004

Maitland, S. B., Intrieri, R. C., Schaie, K. W., & Willis, S. L. (2000). Gender differences and changes in cognitive abilities across the adult life span. *Aging Neuropsychology and Cognition, 7*, 32-53. doi:10.1076/anec.7.1.32.807

Major, J. T., Johnson, W., & Deary, I. J. (2012). Comparing models of intelligence in Project TALENT: The VPR model fits better than the CHC and extended Gf-Gc models. *Intelligence, 40*, 543-559. doi:10.1016/j.intell.2012.07.006

Marsh, H. W. (1989). Confirmatory factor-analyses of multitrait-multimethod data: Many problems and a few solutions. *Applied Psychological Measurement, 13,* 335-361. doi:10.1177/014662168901300402

Marsh, H. W., & Hocevar, D. (1988). A new, more powerful approach to multitrait-multimethod analyses: Application of second-order confirmatory factor analysis. *Journal of Applied Psychology, 73,* 107-117. doi:10.1037/0021-9010.73.1.107

**Martínez, K., Burgaleta, M., Román, F. J., Escorial, S., Shih, P. C., Quiroga, M. Á., & Colom, R. (2011). Can fluid intelligence be reduced to ‘simple’ short-term storage? *Intelligence, 39*, 473-480. doi:10.1016/j.intell.2011.09.001

**Martínez, K., & Colom, R. (2009). Working memory capacity and processing efficiency predict fluid but not crystallized and spatial intelligence: Evidence supporting the neural noise hypothesis. *Personality and Individual Differences, 46*, 281-286. doi:10.1016/j.paid.2008.10.012

McGrew, K. S. (1997). Analysis of the major intelligence batteries according to a proposed comprehensive Gf-Gc framework. In D. P. Flanagan, J. L. Genshaft,

& P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 151-179). New York, NY: Guilford Press.

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence, 37*, 1-10. doi:10.1016/j.intell.2008.08.004

McGrew, K. S., & Flanagan, D. P. (1998). *The intelligence test desk reference (ITDR): Gf-Gc cross-battery assessment*. Boston: Allyn and Bacon.

McInnes, A. J. (2011). *Crystallised and fluid ability change across age and a psychometric evaluation of the GRT2: A cross-sectional analysis* (Unpublished master's thesis). Massey University, Auckland, New Zealand.

Meo, M., Roberts, M. J., & Marucci, F. S. (2007). Element salience as a predictor of item difficulty for Raven's Progressive Matrices. *Intelligence, 35*, 359-368. doi:10.1016/j.intell.2006.10.001

Miller, M. B., Donovan, C. L., Bennett, C. M., Aminoff, E. M., & Mayer, R. E. (2012). Individual differences in cognitive style and strategy predict similarities in the patterns of brain activity between individuals. *NeuroImage, 59*, 83-93. doi:10.1016/j.neuroimage.2011.05.060

Mindrila, D. (2010). Maximum likelihood (ML) and diagonally weighted least squares (DWLS) estimation procedures: A comparison of estimation bias with ordinal and multivariate non-normal data. *International Journal of Digital Society, 1*, 60-66.

Murphy, L. L., Spies, R. A., & Plake, B. S. (2006). *Tests in print VII: An index to tests, test reviews, and the literature on specific tests*. Lincoln, Nebraska: Buros Institute of Mental Measurements.

Muthén, B., & Asparouhov, T. (2002). *Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus*. Retrieved from Mplus website: https://www.statmodel.com/download/webnotes/CatMG Long.pdf

Muthén, B., du Toit, S., & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes*. Retrieved from Mplus website: https://www.statmodel.com/download/Article_075.pdf

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

*Naderi, H., Abdullah, R., Aizan, H. T., & Sharir, J. (2010). Intelligence and academic achievement: An investigation of gender differences. *Life Science Journal-Acta Zhengzhou University Overseas Edition, 7,* 83-87. doi:10.7537/marslsj070110.15

Neisser, U., Boodoo, G., Bouchard, T. J., Boykin, A. W., Brody, N., Ceci, S. J., . . . Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51*, 77-101. doi:10.1037/0003-066x.51.2.77

Njemanze, P. C. (2005). Cerebral lateralization and general intelligence: Gender differences in a transcranial Doppler study. *Brain and Language, 92*, 234-239. doi:10.1016/j.bandl.2004.06.104

Nyborg, H. (2005). Sex-related differences in general intelligence g, brain size, and social status. *Personality and Individual Differences, 39*, 497-509. doi:10.1016/j.paid.2004.12.011

Oakland, T. (1995). 44 country survey shows international test use patterns. *Psychology International, 6*, 7.

Oakland, T., Douglas, S., & Kane, H. (2016). Top ten standardized tests used internationally with children and youth by school psychologists in 64 countries: A 24-year follow-up study. *Journal of Psychoeducational Assessment, 34*, 166-176. doi:10.1177/0734282915595303

Odendaal, A. (2015). Cross-cultural differences in social desirability scales: Influence of cognitive ability. *South African Journal of Industrial Psychology, 41,* 1-13. doi:10.4102/sajip.v41i1.1259

Ortiz, S. O. (2015). CHC theory of intelligence. In S. Goldstein, D. Princiotta, & J. A. Naglieri (Eds.)*, Handbook of intelligence: Evolutionary theory, hisotrical perspective, and current concepts* (pp. 209-227). New York, NY: Springer Science.

Owen, K. (1992). The suitability of Raven's Standard Progressive Matrices for various groups in South Africa. *Personality and Individual Differences, 13*, 149-159. doi:10.1016/0191-8869(92)90037-P

Ozer, D. J. (1987). Personality, intelligence, and spatial visualization: correlates of mental rotations test performance. *Journal of Personality and Social Psychology, 53*, 129-134. doi:10.1037/0022-3514.53.1.129

*Parker, E. S., Parker, D. A., & Harford, T. C. (1991). Specifying the relationship between alcohol-use and cognitive loss: The effects of frequency of consumption and psychological distress. *Journal of Studies on Alcohol, 52*, 366-373. doi:10.15288/jsa.1991.52.366

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. *The Journal of Experimental Education, 54,* 95-100. doi:10.1080/00220973.1986.10806404

*Piffer, D. (2016). Sex differences in intelligence on the American WAIS-IV. *Mankind Quarterly, 57*, 25-33. Retrieved from http://mankindquarterly.org/

Pind, J., Gunnarsdottir, E. K., & Johannesson, H. S. (2003). Raven's Standard Progressive Matrices: New school age norms and a study of the test's validity. *Personality and Individual Differences, 34*, 375-386. doi:10.1016/S0191-8869(02)00058-2

Plaisted, K., Bell, S., & Mackintosh, N. J. (2011). The role of mathematical skill in sex differences on Raven's Matrices. *Personality and Individual Differences, 51,* 562-565. doi:10.1016/j.paid.2011.05.005

*Ponton, M. O., Satz, P., Herrera, L., Ortiz, F., Urrutia, C. P., Young, R., . . . Namerow, N. (1996). Normative data stratified by age and education for the Neuropsychological Screening Battery for Hispanics (NeSBHIS): Initial report. *Journal of the International Neuropsychological Society, 2*, 96-104. doi:10.1017/S1355617700000941

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychology Bulletin, 135*, 322-338. doi:10.1037/a0014996

Postlethwaite, B. E. (2011). *Fluid ability, crystallized ability, and performance across multiple domains: A meta-analysis* (Unpublished doctoral dissertation). University of Iowa.

Prabhakaran, V., Smith, J. A., Desmond, J. E., Glover, G. H., & Gabrieli, J. D. (1997). Neural substrates of fluid reasoning: An fMRI study of neocortical activation during performance of the Raven's Progressive Matrices Test. *Cognitive Psychology, 33*, 43-63. doi:10.1006/cogp.1997.0659

Primi, R. (2001). Complexity of geometric inductive reasoning tasks: Contribution to the understanding of fluid intelligence. *Intelligence, 30*, 41-70. doi:10.1016/S0160-2896(01)00067-8

Primi, R. (2014). Developing a fluid intelligence scale through a combination of Rasch modeling and cognitive psychology. *Psychological Assessment, 26*, 774-788. doi:10.1037/a0036712

Psytech International. (n.d.). *General and graduate reasoning tests*. Bedfordshire, UK: Author.

Quereshi, M. Y., & Seitz, R. (1993). Gender differences in reasoning ability measured by letter series items. *Current Psychology, 12*, 268-272. doi:10.1007/Bf02686808

**Quiroga, M., Escorial, S., Román, F. J., Morillo, D., Jarabo, A., Privado, J., . . . Colom, R. (2015). Can we reliably measure the general factor of intelligence (g) through commercial video games? Yes, we can! *Intelligence, 53*, 1-7. doi:10.1016/j.intell.2015.08.004

R Core Team (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Raabe, S., Hoger, R., & Delius, J. D. (2006). Sex differences in mental rotation strategy. *Perceptual and Motor Skills, 103*, 917-930. doi:10.2466/pms.103.3.917-930

Raîche, G. (2005). Critical eigenvalue sizes in standardized residual principal components analysis. *Rasch Measurement Transactions, 19*, 1012. Retrieved from http://www.rasch.org/rmt/rmt191h.htm

*Rammsayer, T., & Troche, S. (2010). Sex differences in the processing of temporal information in the sub-second range. *Personality and Individual Differences, 49*, 923-927. doi:10.1016/j.paid.2010.07.031

Raven, J. (1962). *Advanced Progressive Matrices.* Oxford, England: Oxford Psychologists Press.

Raven, J. (2008). General introduction and overview: The Raven Progressive Matrices Tests: Their theoretical basis and measurement model. In J. Raven & J. Raven (Eds.), *Uses and abuses of intelligence: Studies advancing Spearman and Raven's quest for non-arbitrary metrics* (pp. 17-68). Unionville, New York: Royal Fireworks Press.

Raven, J., Raven, J. C., & Court, J. H. (1998). *Manual for Raven's progressive matrices and vocabulary scales. Section 3: Standard progressive matrices*.

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

*Raz, N., Rodrigue, K. M., Kennedy, K. M., & Land, S. (2009). Genetic and vascular modifiers of age-sensitive cognitive skills: Effects of COMT, BDNF, ApoE, and hypertension. *Neuropsychology, 23,* 105-116. doi:10.1037/a0013487

Ren, X., Goldhammer, F., Moosbrugger, H., & Schweizer, K. (2012). How does attention relate to the ability-specific and position-specific components of reasoning measured by APM? *Learning and Individual Differences, 22,* 1-7. doi:10.1016/j.lindif.2011.09.009

Ren, X., Schweizer, K., Wang, T., & Xu, F. (2015). The prediction of students' academic performance with fluid intelligence in giving special consideration to the contribution of learning. *Advances in Cognitive Psychology, 11*, 97-105. doi:10.5709/acp-0175-z

Ren, X. Z., Wang, T. F., Altmeyer, M., & Schweizer, K. (2014). A learning-based account of fluid intelligence from the perspective of the position effect. *Learning and Individual Differences, 31*, 30-35. doi:10.1016/j.lindif.2014.01.002

Revelle, W. (2015) psych: Procedures for Personality and Psychological Research (R Package version 1.5.8), Northwestern University, Evanston, Illinois, USA. Retrieved from: http://CRAN.R-project.org/package=psych

Reynolds, M. R., Hajovsky, D. B., Niileksela, C. R., & Keith, T. Z. (2011). Spearman's law of diminishing returns and the DAS-II: Do g effects on subtest scores depend on the level of g? *School Psychology Quarterly, 26*, 275-289. doi:10.1037/a0026190

Reynolds, M. R., Keith, T. Z., Ridley, K. P., & Patel, P. G. (2008). Sex differences in latent general and broad cognitive abilities for children and youth: Evidence from higher-order MG-MACS and MIMIC models. *Intelligence, 36*, 236-260. doi:10.1016/j.intell.2007.06.003

Rindermann, H., & Neubauer, A. C. (2004). Processing speed, intelligence, creativity, and school performance: Testing of causal hypotheses using structural equation models. *Intelligence, 32*, 573-589. doi:10.1016/j.intell.2004.06.005

Roberts, M. J., Welfare, H., Livermore, D. P., & Theadom, A. M. (2000). Context, visual salience, and inductive reasoning. *Thinking and Reasoning*, *6*, 349-374. doi:10.1080/135467800750038175

Roberts, R. D., & Stankov, L. (1999). Individual differences in speed of mental processing and human cognitive abilities: Toward a taxonomic model. *Learning and Individual Differences*, *11*, 1-120. doi:10.1016/S1041-6080(00)80007-2

Rohde, T. E., & Thompson, L. A. (2007). Predicting academic achievement with cognitive ability. *Intelligence, 35*, 83-92. doi:10.1016/j.intell.2006.05.004

Rosen, M. (1995). Gender differences in structure, means and variances of hierarchically ordered ability dimensions. *Learning and Instruction, 5*, 37-62. doi:10.1016/0959-4752(95)00002-K

Roth, B., Becker, N., Romeyke, S., Schafer, S., Domnick, F., & Spinath, F. M. (2015). Intelligence and school grades: A meta-analysis. *Intelligence, 53*, 118-137. doi:10.1016/j.intell.2015.09.002

*Rushton, J. P., & Čvorović, J. (2009). Data on the Raven's Standard Progressive Matrices from four Serbian samples. *Personality and Individual Differences*, *46*, 483-486. doi:10.1016/j.paid.2008.11.020

*Saccuzzo, D. P., Craig, A. S., Johnson, N. E., & Larson, G. E. (1996). Gender differences in dynamic spatial abilities. *Personality and Individual Differences, 21*, 599-607. doi:10.1016/0191-8869(96)00090-6

Salthouse, T. (2012). Consequences of age-related cognitive declines. *Annual Review of Psychology*, *63*, 201-226. doi:10.1146/annurev-psych-120710-100328

Salthouse, T. A. (2005a). Effects of aging on reasoning. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 589-605). New York, NY: Cambridge University Press.

Salthouse, T. A. (2005b). Relations between cognitive abilities and measures of executive functioning. *Neuropsychology, 19*, 532-545. doi:10.1037/0894-4105.19.4.532

Salthouse, T. A. (2016). Aging cognition unconfounded by prior test experience. *The Journals of Gerontology Series B Psychological Sciences and Social Sciences, 71,* 49-58. doi:10.1093/geronb/gbu063

*Salthouse, T. A., & Mitchell, D. R. D. (1990). Effects of age and naturally-occurring experience on spatial visualization performance. *Developmental Psychology, 26,* 845-854. doi:10.1037//0012-1649.26.5.845

Salthouse, T. A., Pink, J. E., & Tucker-Drob, E. M. (2008). Contextual analysis of fluid intelligence. *Intelligence, 36*, 464-486. doi:10.1016/j.intell.2007.10.003

Salzberger, T. (2002). The illusion of measurement: Rasch versus 2-PL. *Rasch Measurement Transactions, 16*, 882. Retrieved from https://www.rasch.org/rmt/rmt162j.htm

Savage-McGlynn, E. (2012). Sex differences in intelligence in younger and older participants of the Raven's Standard Progressive Matrices Plus. *Personality and Individual Differences, 53,* 137-141. doi:10.1016/j.paid.2011.06.013

*Schaie, K. W., Caskie, G. I., Revell, A. J., Willis, S. L., Kaszniak, A. W., & Teri, L. (2005). Extending neuropsychological assessments into the primary mental ability space. *Neuropsychology, Development and Cognition, 12*, 245-277. doi:10.1080/13825580590969343

Schmidt, F. L., & Hunter, J. (2004). General mental ability in the world of work: Occupational attainment and job performance. *Journal of Personality and Social Psychology, 86,* 162-173. doi:10.1037/0022-3514.86.1.162

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124,* 262-274. doi:10.1037/0033-2909.124.2.262

Schneider, W. J., & McGrew, K. S. (2012). The Cattell-Horn-Carroll model of intelligence. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, test, and issues* (3rd ed., pp. 99-144). New York: Guilford.

Schweizer, K. (2012). The position effect in reasoning items considered from the CFA perspective. *The International Journal of Educational and Psychological Assessment, 11*, 44-58. Retrieved from https://sites.google.com/site/tijepa2012/home

Schweizer, K., Goldhammer, F., Rauch, W., & Moosbrugger, H. (2007). On the validity of Raven’s matrices test: Does spatial ability contribute to performance? *Personality and Individual Differences, 43*, 1998-2010. doi:10.1016/j.paid.2007.06.008

Schweizer, K., Reiss, S., Schreiner, M., & Altmeyer, M. (2012). Validity improvement in two reasoning measures following the elimination of the position effect. *Journal of Individual Differences, 33*, 54-61. doi:10.1027/1614-0001/a000062

Schweizer, K., Schreiner, M., & Gold, A. (2009). The confirmatory investigation of APM items with loadings as a function of the position and easiness of items: A two-dimensional model of APM. *Psychology Science, 51*, 47-64. Retrieved from http://www.psychologie-aktuell.com/index.php?id=200

Schweizer, K., Troche, S. J., & Rammsayer, T. H. (2011). On the special relationship between fluid and general intelligence: New evidence obtained by considering the position effect. *Personality and Individual Differences, 50*, 1249-1254. doi:10.1016/j.paid.2011.02.019

*Sellami, K., Infanzon, E., Lanzon, T., Diaz, A., & Lynn, R. (2010). Sex Differences in means and variance of intelligence: Some data from Morocco. *Mankind Quarterly, 51*, 84-92. Retrieved from http://mankindquarterly.org/

Sick, J. (2010). Assumptions and requirements of Rasch measurement. *SHIKEN: JALT Testing and Evaluation SIG Newsletter, 14,* 23-29. Retrieved from http://jalt.org/test/PDF/Sick5.pdf

Sieff, M., & Uttal, D. (2015). How much can spatial training improve STEM achievement? *Educational Psychology Review, 27,* 607-615. doi:10.1007/s10648-015-9304-8

*Silvia, P. J., & Sanders, C. E. (2010). Why are smart people curious? Fluid intelligence, openness to experience, and interest. *Learning and Individual Differences, 20,* 242-245. doi:10.1016/j.lindif.2010.01.006

Soares, D. L., Lemos, G. C., Primi, R., & Almeida, L. S. (2015). The relationship between intelligence and academic achievement throughout middle school: The role of students' prior academic performance. *Learning and Individual Differences, 41,* 73-78. doi:10.1016/j.lindif.2015.02.005

Spearman, C. (1923). *The nature of "intelligence" and the principles of cognition*. London, England: MacMillan.

Spearman, C. (1927). *The abilities of man*. London: MacMillan.

Stankov, L. (1997). *The Gf/Gc Quickie Test Battery* (Unpublished test battery available from the School of Psychology). University of Sydney.

Stankov, L., & Cregan, A. (1993). Quantitative and qualitative properties of an intelligence test: Series completion. *Learning and Individual Differences, 5,* 137-169. doi:10.1016/1041-6080(93)90009-H

*Stein, A. D., Behrman, J. R., DiGirolamo, A., Grajeda, R., Martorell, R., Quisumbing, A., & Ramakrishnan, U. (2005). Schooling, educational achievement, and cognitive functioning among young Guatemalan adults. *Food and Nutrition Bulletin, 26,* S46-S54.

Steinmayr, R., Beauducel, A., & Spinath, B. (2010). Do sex differences in a faceted model of fluid and crystallized intelligence depend on the method applied? *Intelligence, 38,* 101-110. doi:10.1016/j.intell.2009.08.001

Stephenson, C. L., & Halpern, D. F. (2013). Improved matrix reasoning is limited to training on tasks with a visuospatial component. *Intelligence, 41,* 341-357. doi:10.1016/j.intell.2013.05.006

*Stewart, M. C., Deary, I. J., Fowkes, F. G., & Price, J. F. (2006). Relationship between lifetime smoking, smoking status at older age and human cognitive function. *Neuroepidemiology, 26,* 83-92. doi:10.1159/000090253

Strand, S., Deary, I. J., & Smith, P. (2006). Sex differences in Cognitive Abilities Test scores: A UK national picture. *British Journal of Educational Psychology, 76,* 463-480. doi:10.1348/000709905X50906

Süß, H. M., & Beauducel, A. (2005). Faceted models of intelligence. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 313-332). London, UK: Sage.

Süß, H. M., & Beauducel, A. (2015). Modeling the construct validity of the Berlin Intelligence Structure Model. *Estudos de Psicologia, 32,* 13-25. doi:10.1590/0103-166X2015000100002

*Tan, U. (1991). The inverse relationship between nonverbal intelligence and the latency of the Hoffmann reflex from the right and left Thenar muscles in righthanded and left-handed subjects. *International Journal of Neuroscience, 57,* 219-238. doi:10.3109/00207459109150696

Thurstone, L. L. (1938). *Primary mental abilities*. Chicago: University of Chicago Press.

**Tommasi, M., Pezzuti, L., Colom, R., Abad, F. J., Saggino, A., & Orsini, A. (2015). Increased educational level is related with higher IQ scores but lower g-variance: Evidence from the standardization of the WAIS-R for Italy. *Intelligence, 50*, 68-74. doi:10.1016/j.intell.2015.02.005

*Tuttle, G. E., & Pillard, R. C. (1991). Sexual orientation and cognitive abilities. *Archives of Sexual Behavior, 20*, 307-318. doi:10.1007/BF01541849

Valla, J. M., & Ceci, S. J. (2014). Breadth-based models of women's underrepresentation in STEM fields: An integrative commentary on Schmidt (2011) and Nye et al. (2012). *Perspectives on Psychological Science, 9*, 219-224. doi:10.1177/1745691614522067

Van der Elst, W., Ouwehand, C., van Rijn, P., Lee, N., Van Boxtel, M., & Jolles, J. (2013). The shortened Raven Standard Progressive Matrices: Item response theory-based psychometric analyses and normative data. *Assessment, 20*, 48-59. doi:10.1177/1073191111415999

*van der Sluis, S., Posthuma, D., Dolan, C. V., de Geus, E. J. C., Colom, R., & Boomsma, D. I. (2006). Sex differences on the Dutch WAIS-III. *Intelligence, 34*, 273-289. doi:10.1016/j.intell.2005.08.002

van der Ven, A. H. G. S., & Ellis, J. L. (2000). A Rasch analysis of Raven's Standard Progressive Matrices. *Personality and Individual Differences, 29*, 45-64. doi:10.1016/S0191-8869(99)00177-4 *van Leeuwen, M., van den Berg, S. M., & Boomsma, D. I. (2008). A twin-family study of general IQ. *Learning and Individual Differences, 18*, 76-88. doi:10.1016/j.lindif.2007.04.006

Vandenberg, S. G., & Kuse, A. R. (1978). Mental rotation: A group test of three-dimensional spatial visualisation. *Perceptual and Motor Skills, 47,* 599-604. doi:10.2466/pms.1978.47.2.599

Vernon, P. E. (1964). *The structure of human abilities*. London: Methuen.

Viechtbauer, W. (2010). Conducting meta-analysis in R with the metaphor package. *Journal of Statistical Software, 36,* 1-48. doi:10.18637/jss.v036.i03

*Vigneau, F., & Bors, D. A. (2005). Items in context: Assessing the dimensionality of Raven's Advanced Progressive Matrices. *Educational and Psychological Measurement, 65*, 109-123. doi:10.1177/0013164404267286

*Vigneau, F., & Bors, D. A. (2008). The quest for item types based on information processing: An analysis of Raven's Advanced Progressive Matrices, with a consideration of gender differences. *Intelligence, 36*, 702-710. doi:10.1016/j.intell.2008.04.004

*von Stumm, S., Chamorro-Premuzic, T., Quiroga, M. A., & Colom, R. (2009). Separating narrow and general variances in intelligence-personality associations. *Personality and Individual Differences, 47*, 336-341. doi:10.1016/j.paid.2009.03.024

Voyer, D., Voyer, S., & Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. *Psychological Bulletin, 117,* 250-270. doi:10.1037/0033-2909.117.2.250

Voyer, D., & Voyer, S. D. (2014). Gender differences in scholastic achievement: A meta-analysis. *Psychological Bulletin, 140,* 1174-1204. doi:10.1037/a0036620

*Wachs, T. D., McCabe, G., Moussa, W., Yunis, F., Kirksey, A., Galal, O., . . . Jerome, N. (1996). Cognitive performance of Egyptian adults as a function of nutritional intake and sociodemographic factors. *Intelligence, 22*, 129-154. doi:10.1016/S0160-2896(96)90013-6

Wai, J., Lubinski, D., & Benbow, C. P. (2009). Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. *Journal of Educational Psychology, 101*, 817-835. doi:10.1037/a0016127

Wang, L., & Carr, M. (2014). Working memory and strategy use contribute to gender differences in spatial ability. *Educational Psychologist, 49*, 261-282. doi:10.1080/00461520.2014.960568

*Welborn, B. L., Papademetris, X., Reis, D. L., Rajeevan, N., Bloise, S. M., & Gray, J. R. (2009). Variation in orbitofrontal cortex volume: Relation to sex, emotion regulation and affect. *Social Cognitive and Affective Neuroscience, 4*, 328-339. doi:10.1093/scan/nsp028

Weng, L. J., & Cheng, C. P. (2005). Parallel analysis with unidimensional binary data. *Educational and Psychological Measurement, 65*, 791-810. doi:10.1177/0013164404273941

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement, 9*, 1-26. doi:10.1177/014662168500900101

Wilhelm, O. (2005). Measuring reasoning ability. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence*. California: Sage.

Yang, W., Liu, P., Wei, D., Li, W., Hitchman, G., Li, X., . . . Zhang, Q. (2014). Females and males rely on different cortical regions in Raven's Matrices reasoning capacity: Evidence from a voxel-based morphometry study. *PLoS One, 9*, e93104. doi:10.1371/journal.pone.0093104

Yoon, S. Y. (2011). *Psychometric properties of the Revised Purdue Spatial Visualization Tests: Visualization of Rotations (The Revised PSVT:R)* (Unpublished doctoral dissertation). Purdue University, West Lafayette, Indiana.

Yu, C. Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes* (Unpublished doctoral dissertation). University of California, Los Angeles.

Zeller, F., Wang, T., Reiß, S., & Schweizer, K. (2017). Does the modality of measures influence the relationship among working memory, learning and fluid intelligence? *Personality and Individual Differences, 105*, 275-279. doi:10.1016/j.paid.2016.10.013

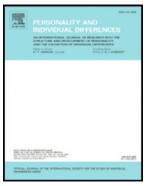
# Appendix A

Personality and Individual Differences 100 (2016) 157–166

Contents lists available at ScienceDirect

Personality and Individual Differences

journal homepage: www.elsevier.com/locate/paid



## Dimensionality of the Raven's Advanced Progressive Matrices: Sex differences and visuospatial ability☆

Nicolette A. Waschl [a,*], Ted Nettelbeck [a], Simon A. Jackson [b], Nicholas R. Burns [a]

[a] *School of Psychology, University of Adelaide, South Australia 5005, Australia*
[b] *School of Psychology, University of Sydney, New South Wales 2006, Australia*

ARTICLE INFO

*Article history:*
Received 28 March 2015
Received in revised form 24 November 2015
Accepted 2 December 2015
Available online 15 December 2015

*Keywords:*
Raven's Progressive Matrices,
Sex differences
Confirmatory factor analysis
Rasch analysis

ABSTRACT

Raven's progressive matrices are considered a measure of inductive reasoning. However, there is evidence to suggest that they are not unidimensional, and they may measure visuospatial ability in addition to inductive reasoning. We investigated the psychometric properties of several versions of the Advanced Progressive Matrices (APM). Confirmatory factor analyses and Rasch analyses were used to investigate the dimensionality of the test, sex differences regarding dimensionality, and the utility of proposed taxonomies of item solution strategies. Three samples were administered three different forms of the test. Sample 1 consisted of 1297 individuals (929 females) who completed a 12-item short form; Sample 2 consisted of 455 individuals (327 females) who completed the full APM; and Sample 3 consisted of 362 individuals (244 females) who completed a 15-item short form. Results indicated that all three forms of the APM are unidimensional and measurement invariant across sex. There was little support for the validity of the taxonomies of solution strategies.

Raven's progressive matrices (RPM) were designed to measure Spearman's *g*. Under the Cattell–Horn–Carroll model of intelligence (McGrew, 2009), the RPM tests (including the coloured, standard and advanced versions, designed for use with different populations) measure fluid intelligence and, specifically, inductive reasoning. However, there has been speculation that they also measure visuospatial ability (see Burke, 1958, for an early review). Fluid ability involves solving unfamiliar problems, while inductive reasoning, a narrow ability under fluid ability, involves discovering underlying principles or rules (McGrew, 2009). Visuospatial ability is different. It involves perceiving, generating and operating on visual patterns and stimuli, and is typified by tasks requiring perception and manipulation of visual forms (McGrew, 2009). It is clear why the claim that the RPM involves visuospatial ability emerged; RPM items comprise visual stimuli and it is conceivable that solving items could require visual transformation of these stimuli. This question was posed more than half a century ago, yet it remains unresolved. This paper focuses on the Advanced Progressive Matrices (APM).

The claim that the APM involves visuospatial ability has important implications. It is essential to understand what such a commonly used and potentially high-stakes test measures in order to understand how scores can be interpreted, used appropriately, and related to other constructs. Additionally, there is evidence of sex differences, favouring males, in APM performance (Lynn & Irwing, 2004a, 2004b). One of the most robust findings in the literature is a male advantage on visuospatial ability tests, particularly mental rotation (Linn & Petersen, 1985; Voyer, Voyer & Bryden, 1995). Therefore, one explanation, other than in terms of inductive reasoning, of a male advantage on the APM, could be the contribution of visuospatial ability to performance. Indeed, there is evidence that visuospatial ability accounts for the observed sex difference in APM scores (Colom, Escorial, & Rebollo, 2004). Through an understanding of whether the APM is unidimensional or multidimensional, and if this differs in male and females, we can come closer to understanding if and how visuospatial ability is involved.

Three different strategies have been used to understand what construct or constructs the APM measures: creation and examination of solution taxonomies based on information processing theories; investigation of sex differences in relation to these taxonomies; and factor analysis. This paper expands on these methods using three different versions of the APM.

Concerning solution taxonomies, Carpenter, Just, and Shell (1990) used patterns of eye fixations and verbalization of solution strategies to determine how each item was solved, resulting in a taxonomy of five solution rules (Table 1). These rules were: constant in a row; quantitative pairwise progression; addition/subtraction; distribution of three; and distribution of two. Constant in a row is not considered further because it always occurs in conjunction with another rule. Quantitative pairwise progression involves a quantitative increment or decrement across the row in size, position or number; Addition/Subtraction involves adding or subtracting a figure in one column from another figure to produce the third; Distribution of three is when

☆ This article is a Special issue article – “Young researcher award 2015".
* Corresponding author.
*E-mail addresses:* nicolette.waschl@adelaide.edu.au (N.A. Waschl), theodore.nettelbeck@adelaide.edu.au (T. Nettelbeck), simon.jackson@sydney.edu.au (S.A. Jackson), nicholas.burns@adelaide.edu.au (N.R. Burns).

http://dx.doi.org/10.1016/j.paid.2015.12.008
0191-8869/© 2015 Elsevier Ltd. All rights reserved.

**Table 1**
Classifications of APM Items.

| Item | Carpenter et al. (1990) | DeShon et al. (1995) | Dillon et al. (1981) | Item | Carpenter et al. (1990) | DeShon et al. (1995) | Dillon et al. (1981) |
|---|---|---|---|---|---|---|---|
| 1 | D3 | Analytic | | 19 | A/S | Both | |
| 2 | | Either | PP | 20 | A/S | Both | |
| 3 | P | Visual | PP | 21 | D3 | Analytic | A/S |
| 4 | P | Analytic | PP | 22 | D2 | Visual | |
| 5 | P | Either | PP | 23 | D2 | Visual | |
| 6 | P | Either | | 24 | P | Visual | |
| 7 | A/S | Visual | A/S | 25 | P | Both | |
| 8 | D3 | Analytic | | 26 | P, D3 | Both | PP |
| 9 | A/S | Visual | A/S | 27 | D3 | Analytic | |
| 10 | P | Visual | A/S | 28 | D3 | Analytic | A/S |
| 11 | | Visual | A/S | 29 | D3 | Analytic | |
| 12 | A/S | Visual | | 30 | D2 | Analytic | |
| 13 | D3 | Analytic | | 31 | D3, D2 | Both | |
| 14 | P | Either | | 32 | D3, D2 | Visual | |
| 15 | A/S | | | 33 | A/S | Visual | |
| 16 | A/S | Visual | A/S | 34 | D3 | Analytic | |
| 17 | D3 | Analytic | PP | 35 | D2 | Both | A/S |
| 18 | | Visual | | 36 | D2 | Analytic | PP |

*Note*. P = quantitative pairwise progression; A/S = addition/subtraction; D3 = distribution of 3; D2 = distribution of 2; PP = pattern progression; both = analytic and visual; either = analytic or visual. Carpenter et al.'s (1990) classifications are supplemented by Mackintosh and Bennett (2005).

three values from a categorical attribute are distributed across the row; and Distribution of two is when two values from a categorical attribute are distributed across a row and the third value is null.

Following Carpenter et al. (1990), DeShon, Chan, and Weissbein (1995) expanded on these rules and obtained 12 solution rules; four involved verbal-analytic processes and eight involved visual processes. Although these taxonomies are not directly comparable, Carpenter et al.'s addition/subtraction rule tended to equate with DeShon et al.'s visual process, while distribution of three tended to equate with an analytic process.

Given the well-established male advantage on visuospatial ability tests, and the grouping of solution rules into verbal-analytic and visual types, these solution taxonomies have been studied in relation to sex differences on the APM. Mackintosh and Bennett (2005) found a male advantage on items involving addition/subtraction and distribution of two, argued to involve visual processes, but no sex difference in items involving quantitative pairwise progression and distribution of three, argued to involve analytic processes. Other studies, however, have found no consistent sex differences in these item types (Vigneau & Bors, 2008), or a male advantage on all types (Colom & Abad, 2007). The picture of how these item types may relate to sex differences in scores is not clear.

Similarly, while factor analysis has commonly been used to investigate the structure of the APM, it has yet to provide a solution to the question of dimensionality, or the role of visuospatial ability in performance. One of the most cited factor analyses of the APM was by Dillon, Pohlmann, and Lohman (1981). Using a principal components analysis of phi/phi(max) coefficients, these authors reported two orthogonal factors, pattern progression and addition/subtraction (see Table 1). Addition/subtraction is broadly similar to Carpenter et al.'s (1990) addition/subtraction (although it is represented by different items in Dillon et al.'s study); while pattern progression involves perceiving a recurring or sequential design. However, later research has not supported Dillon et al.'s factors (Alderton & Larson, 1990; Arthur, Tubre, Paul, & Sanchez-Ku, 1999; Arthur & Woehr, 1993; Bors & Stokes, 1998; Vigneau & Bors, 2008) and other factor analyses have tended to indicate a single-factor structure (Abad, Colom, Rebollo, & Escorial, 2004; Chiesi, Ciancaleoni, Galli, Morsanyi & Primi, 2012a; Schweizer, Goldhammer, Rauch, & Moosbrugger, 2007).

Both Carpenter et al. (1990) and DeShon et al.'s (1995) taxonomies have been used in factor analytic studies investigating the dimensionality of the APM. Unfortunately, although these rules have been useful in understanding the cognitive processing strategies that individuals use in solving individual items, there is little support for the idea that these rules represent different latent factors or relate to different latent abilities (Vigneau & Bors, 2008). One aspect yet to be investigated in relation to these solution taxonomies and sex differences in APM performance, however, is whether the latent structure of this test differs across sexes. There is some evidence that it may. For example, Lim (1994) found that the APM loaded on only one factor, formal operations, in males, but two, formal operations and spatial, in females. If this were the case, it could explain some of the inconsistent findings regarding the factor structure of the test. Relatedly, whether or not the test is measurement invariant across sex is important when considering the possibility of different factor structures among males and females, and when considering sex differences in the underlying construct. Despite consideration of the role of visuospatial ability and sex differences in APM performance, little concern has been given to establishing measurement invariance across sex in this test.

Although factor analytic studies have largely supported a unidimensional conceptualization of the APM, other lines of evidence indicate that the APM contains a visuospatial component or, at least, is not unidimensional. This evidence comes from studies using statistical control of visuospatial ability (Colom et al., 2004), experimental manipulation (DeShon et al., 1995), item response theory analysis (Vigneau & Bors, 2005) and neuroimaging (Ebisch et al., 2012); this uncertainty indicates that the matter deserves further consideration. The common finding of unidimensionality in the APM may be partially due to the various issues inherent in the use of factor analysis to answer this question. Ordinary factor analytic methods (e.g. principal axis factoring, maximum likelihood) applied to binary data can be problematic (Hattie, 1985). On the other hand, the weighted least squares mean and variance adjusted (WLSMV) estimator has several advantages over other methods. It was designed specifically for use with binary data and simulation studies have shown it to be appropriate for these types of data (Muthén, du Toit, & Spisic, 1997). However, this estimation method has not yet been applied to the full form APM or the two short forms considered in the present study (Bors & Stokes, 1998; and a form unique to this study).

Another method for investigating the dimensionality of the APM utilizes item response theory (IRT), which is not subject to the same issues as factor analytic methods. Unlike factor analysis, IRT was created for binary data and is therefore appropriate for use with the data obtained from the correct-incorrect responses to APM items. While factor analysis conducted using the WLSMV estimator and IRT are mathematically highly similar, their distinct theoretical standpoints provide an interesting comparison. There are several different IRT models, including the Rasch, 2PL and 3PL models. The Rasch model considers the probability of a correct response to an item given the test-taker's ability and the item difficulty while holding constant item discrimination and guessing. The 2PL and 3PL models allow estimation of other parameters in addition to difficulty; the 2PL model allows estimation of item discrimination, while the 3PL model allows estimation of discrimination and guessing. The Rasch model has excellent measurement properties and well-developed statistical theory, and hence has been used here.

While some studies have used IRT to investigate the APM, few have applied the Rasch model, and those that have did not consider sex differential item functioning (DIF; Vigneau & Bors, 2005). DIF has been considered under the 2PL (Abad et al., 2004) and 3PL models (Chiesi, Ciancaleoni, Galli, Morsanyi & Primi, 2012a; Chiesi, Ciancaleoni, Galli & Primi, 2012b), with conflicting findings. Using the 2PL model, Abad et al. (2004) showed more DIF for items classified as visuospatial than items classified as analytic, while Chiesi, Ciancaleoni, Galli, Morsanyi and Primi (2012a) and Chiesi, Ciancaleoni, Galli and Primi (2012b) work indicated no DIF.

Hence, while there has been a significant amount of work conducted on whether the APM is unidimensional or whether it involves a second ability, hypothesised to be visuospatial ability, questions still remain. The aims of this study were threefold. First, to investigate sex differences in individual items and item types in an attempt to clarify the existing literature and as a prelude to investigating sex differences in the latent structure of the test. Secondly, we applied factor analytic methods not yet used on the APM to test models based on results of Carpenter et al. (1990), DeShon et al. (1995) and Dillon et al. (1981) in males and females separately, and, if appropriate, to examine measurement invariance and latent mean differences. Thirdly, we examined Rasch model fit and DIF to supplement the factor analytic results.

## 1. Method

### 1.1. Participants and measures

All participants provided informed consent before participating in these studies. Participants in Sample 1 were 1297 individuals tested through the University of Adelaide, Australia, most of whom were university students. Participants completed a 12-item short form of the APM (Bors & Stokes, 1998) online and in their own time as part of their coursework. The items originated from Set II of the APM, and were included based on high item-total correlations and low inter-item correlations. The final pool of items consisted of items 3, 10, 12, 15, 16, 18, 21, 22, 28, 30, 31 and 34 from the full form. Nine cases with a score of zero were deleted from this dataset because it was presumed that these participants did not understand the instructions or had not taken the test seriously. The final sample consisted of 1288 (929 females) aged 16 to 60 years ($M = 23.5$, $SD = 6.62$). The mean score for this sample was 7.19 items (60% correct; $SD = 2.71$). Males ($M = 7.53$ [63%], $SD = 2.68$) scored slightly but significantly higher than females ($M = 7.06$ [59%], $SD = 2.71$), $t(1286) = 2.78$, $p = .005$, $d = .17$.

Participants in Sample 2 were 455 adults (327 females) aged 16 to 68 years ($M = 34.47$, $SD = 16.9$) residing in Adelaide and recruited over two studies (see Burns, Bastian & Nettelbeck, 2007). Each participant completed a paper-and-pencil version of Set II of the full form of the APM (36 items; Raven, 1962). The mean score for Sample 2 was 21.34 (59%; $SD = 7.24$). Males ($M = 22.88$ [64%], $SD = 6.79$) scored significantly higher than females ($M = 20.74$ [58%], $SD = 7.34$), $t(453) = 2.845$, $p = .005$, $d = .30$.

Participants in Sample 3 were 362 undergraduate psychology students from the University of Sydney, recruited over two studies (Jackson, Kleitman, Stankov & Howie, n.d.). Participants completed a 15-item short form of the APM, unique to these studies, as part of a larger test battery either online in their own time or as part of their regular tutorial programme. The items originated from Set II of the APM, and the criteria for inclusion of items was based on pilot testing designed to obtain a more pure measure of the APM by selecting those items showing high item-total and inter-item correlations, and high standard deviations. The final pool of items consisted of items 7, 11, 13, 15, 16, 17, 18, 21, 23, 25, 26, 27, 30, 32 and 34 from the full form. Two cases with a score of zero were removed from this dataset and the final sample consisted of 360 (244 females) aged 17 to 54 years ($M = 20.09$, $SD = 3.19$). The mean score for Sample 3 was 8.39 (56%; $SD = 3.42$). Males ($M = 9.02$ [60%], $SD = 3.70$) scored significantly higher than females ($M = 8.09$ [54%], $SD = 3.24$), t(358) = 2.41, $p = .016$, $d = .28$. The sex difference on this short form was substantially larger than the difference in Sample 1, but did show a similar effect size to Sample 2. The method of item selection for the two short forms was different, resulting in only six (40–50%) common items between these short forms, which may have influenced the magnitude of the differences.

### 1.2. Data analysis

#### 1.2.1. Sex differences in items and item types

Analysis of sex differences in individual items and groups of item types was conducted in R (R Development Core Team, 2014). Chi-square tests were used to examine sex differences in individual items and t-tests were used to investigate sex differences in groups of item types as classified by Carpenter et al. (1990), DeShon et al. (1995) and Dillon et al. (1981).

#### 1.2.2. Confirmatory factor analysis

Confirmatory factor analysis was performed in Mplus 7 (Muthén & Muthén, 1998-2012) using the WLSMV estimator. Several models were compared: models based on the item classifications of Carpenter et al. (1990), DeShon et al. (1995) and Dillon et al. (1981), and a model based on item threshold values (difficulty model).[1] These models were compared with reference to the chi-square value, RMSEA and CFI. Guidelines for interpreting these indices recommend the following cut-off values for acceptable fit: normed chi-square (i.e. $\chi^2$/df) < 2 (Kline, 1998), RMSEA < .05 (Browne & Cudeck, 1993) and CFI > .95 (Hu & Bentler, 1999). In addition to inspection of fit indices, where possible, the multi-factor models were compared to their corresponding one-factor models by comparing the chi-square statistics. Because the WLSMV estimator does not follow the chi-square distribution, the DIFFTEST function (Asparouhov & Muthén, 2006) in Mplus was used to test for differences between the models. This function follows a two-step process: first, the less restrictive model is estimated and the derivatives needed for the chi-square difference test are saved. Secondly, the more restrictive model is estimated and the chi-square difference test is calculated using the derivatives from both models (Muthén & Muthén, 1998-2012). Given that not all items were always included in the multi-factor models because some items were not classified under the relevant taxonomies, the corresponding one-factor models only included those items that were in the multi-factor model.

#### 1.2.3. Measurement invariance and latent mean differences

If factor analysis demonstrated that the best fitting model was the same for males and females, multiple-groups confirmatory factor analysis (MGCFA) was carried out in order to determine, firstly, if measurement invariance across sex could be confirmed, and secondly, if there were any differences in the latent mean (or means if considering a multi-factor model) of APM performance. Measurement invariance was tested using mean and covariance structures (MACS) with delta parameterization in Mplus 7.

MGCFA for categorical indicators is somewhat different from MGCFA with continuous indicators. When the data consist of continuous indicators, increasingly restrictive models are tested by constraining equal the factor loadings, intercepts, and then residual variances of factor indicators. With categorical indicators, MGCFA involves testing factor loadings, thresholds and scale factors. Thresholds are tested instead of intercepts and scale factors may be tested instead of residual variances. Categorical MGCFA also involves a comparison of only two models, a less restrictive and a more restrictive, rather than the usual four when dealing with continuous data. The first, less restrictive model allows the thresholds and factor loadings to vary across groups, while scale factors are constrained at one and factor means are constrained at zero in all groups. This less restrictive model is then compared to the more restrictive model in which thresholds and factor loadings are simultaneously held equal across groups and the scale factor is fixed to one and the factor means constrained at zero in the first group but allowed to vary in the other. If there is a significant difference in the fit between these models, this indicates a violation of measurement invariance. To test for the difference in fit, the DIFFTEST function, as explained in Section 2.2.2, was used. If there was no significant difference in model fit, invariance was met. If a significant difference in model fit was found, the modification indices were inspected to determine which

[1] Bi-factor models were also considered, however these models presented estimation problems and demonstrated poor fit when successfully estimated. Therefore this type of model was deemed too complex for the data.

item may be responsible, and the model was re-tested allowing the factor loading and threshold to vary across groups, and constraining that scale factor to one across groups. Once partial invariance was established (i.e. the problem item's measurement parameters were allowed to vary while all others were constrained equal), examination of latent mean differences was conducted by comparing the latent mean in the non-reference group to that of the reference group, which was constrained at zero. A significant value indicated a significant difference in the latent mean across groups (i.e., across sex).

### 1.2.4. Rasch analysis and differential item functioning

Rasch analysis was conducted using ConQuest 3.0.1 (Adams, Wu, & Wilson, 2012). In order to examine whether the data conformed to a Rasch model, and therefore could be considered unidimensional, person fit and item fit were inspected using the mean square statistics. The infit mean square considers the consistency of the item responses to the item characteristic curve (ICC) for each item, with weighted consideration of the responses of those persons close to the 0.5 probability level for that item. Low infit mean square values indicate item redundancy, while high values are more of a threat to unidimensionality because they indicate that the item discriminates poorly. The criterion for person (case) misfit was a standardized outfit mean square > 5, represented by the t-value. A value exceeding this cutoff indicates an erratic response pattern. The criterion for item misfit was a standardized infit mean square (t-value) > 2. While generally the critical value for the unstandardized infit mean square is in the range of 0.77–1.30 (Adams & Khoo, 1993) in large samples a t-value >2 can be within this range. The t-value allows analysis of strict conformity to a Rasch model (perfect model fit) as opposed to whether or not the items are productive for measurement, for which the unstandardized values are more useful (Linacre, 2002). Therefore, for the purposes of investigating the dimensionality of these measures, it was decided that a stricter rather than a more lenient assessment of fit was appropriate and t-values were applied.

In addition to assessment of dimensionality using item and person fit, a principal components analysis (PCA) of the Rasch residuals (residual variance in the data once the variance explained by the Rasch dimension is accounted for) was performed. If the first principal component of the Rasch residuals is above noise level, this indicates the presence of multidimensionality. A simulation based on the number of items and number of cases was performed for each sample to determine at what point the value of the first eigenvalue exceeded random noise (Raîche, 2005). If this value was exceeded, this indicated multidimensionality.

Following examination of dimensionality, DIF across sex was investigated using two methods; the item fit and the item threshold approaches. The item fit approach involved calibration of the Rasch model separately in the two groups and inspection of the item fit t-values in order to determine if the same items showed misfit. If an item displays acceptable fit in the combined group, but shows misfit in only one of the male or female groups, this indicates a biased item that does not discriminate equally across groups (Hungi, 2005). The item threshold approach used the Wald t statistic, calculated on the basis of the values provided by the item-by-sex interaction parameters in ConQuest. The Wald t statistic is calculated by dividing the item's item-by-group interaction parameter by its standard error. Any item with a t-value >2 displays statistically significant DIF.

## 2. Results

### 2.1. Sex differences in items and item types

In each sample there were individual items showing significant sex differences, all of which showed a higher male score (Table 2). Item numbers are presented as their number in the original form of the test (i.e. the first item in Sample 1 is labelled as item 3) so as to facilitate comparison across forms.

**Table 2**
Sex differences in individual items.

| Test | Item | Carpenter et al. (1990) | DeShon et al. (1995) | Dillon et al. (1981) | Chi-square | Phi |
|---|---|---|---|---|---|---|
| Sample 1 | 12 | A/S | Visual | – | 10.36[**] | .09 |
| | 21 | D3 | Analytic | A/S | 5.61[*] | .07 |
| | 22 | D2 | Visual | – | 7.21[**] | .08 |
| | 28 | D3 | Analytic | A/S | 4.13[*] | .06 |
| Sample 2 | 2 | – | Either | PP | 6.32[*] | .12 |
| | 4 | P | Analytic | PP | 10.02[**] | .15 |
| | 9 | A/S | Visual | A/S | 7.41[**] | .13 |
| | 10 | P | Visual | A/S | 4.46[*] | .10 |
| | 13 | D3 | Analytic | – | 5.48[*] | .11 |
| | 21 | D3 | Analytic | A/S | 4.49[*] | .10 |
| | 22 | D2 | Visual | – | 5.71[*] | .11 |
| | 25 | P | Both | – | 6.95[**] | .12 |
| | 26 | P, D3 | Both | PP | 5.49[*] | .11 |
| Sample 3 | 30 | D2 | Analytic | – | 4.08[*] | .11 |
| | 32 | D3, D2 | Visual | – | 9.55[**] | .16 |

*Note.* P = quantitative pairwise progression; A/S = addition/subtraction; D2 = distribution of two; D3 = distribution of three; PP = pattern progression; both = analytic and visual; either = analytic or visual. All significant differences favour males.
[*] $p < .05$.
[**] $p < .01$.

No items were found to show a consistent sex difference across samples, and there was no clear pattern regarding the classifications of the individual items showing sex differences. Similarly, no consistent pattern of sex differences in item types was found (Table 3). In Samples 1 and 2 nearly all groups of items showed a significant sex difference, while in Sample 3 only distribution of two (Carpenter et al., 1990), visual (DeShon et al., 1995) and addition/subtraction (Dillon et al., 1981) showed significant differences.

### 2.2. Confirmatory factor analysis

This section presents the results of the factor analysis in each sample (see supplementary materials for additional information). In all cases, the difficulty model showed an acceptable fit to the data and was statistically significantly better fitting than the corresponding one-factor model (with the exception of the Sample 3 combined and female data). One common issue with the use of factor analysis with the type of data used in the current study is the occurrence of artifactual factors based on item difficulty. The use of the WLSMV estimator should avoid the issue of these artifactual difficulty factors by allowing a non-linear relationship between the item and the factor, the main cause of this problem (Gibson, 1960). However, the difficulty models are statistically

**Table 3**
Sex differences in item types.

| Taxonomy | Classification | *t* | | |
|---|---|---|---|---|
| | | Sample 1 | Sample 2 | Sample 3 |
| Carpenter et al. (1990) | Pairwise progression | 1.25 | 3.11[**] | – |
| | Addition/subtraction | 2.77[**a] | 2.05[*] | 1.04[b] |
| | Distribution of 3 | 2.51[*] | 2.43[*] | 1.29 |
| | Distribution of 2 | 2.28[*] | 2.35[*] | 2.55[*] |
| DeShon et al. (1995) | Visual | 2.30[*] | 2.49[*] | 2.69[**] |
| | Analytic | 2.33[*] | 2.35[*] | 1.69 |
| Dillon et al. (1981) | Pattern Progression | – | 3.60[**] | 0.71[c] |
| | Addition/Subtraction | 2.59[**] | 2.77[**] | 2.01[*] |

*Note.* Sample 1 *df* = 1286; Sample 2 *df* = 453; and Sample 3 *df* = 358. All significant differences favour males.
Blank cells indicate groups that could not be tested because there were not enough items to represent the classification.
[a] *df* = 732.47.
[b] *df* = 283.70.
[c] *df* = 295.54.
[*] $p < .05$.
[**] $p < .01$.

**Table 4**
Model fit indices: Sample 1.

| Model | | $\chi^2$ | *df* | CFI | RMSEA | Factor correlation | |
|---|---|---|---|---|---|---|---|
| | | | | | | Estimate | 95% CI |
| *One-factor* | | | | | | | |
| | Combined | 103.84 | 54 | .98 | .03 | | |
| | Female | 84.60 | 54 | .98 | .03 | | |
| | Male | 80.32 | 54 | .97 | .04 | | |
| *Two factors: DeShon et al. (1995)* | | | | | | | |
| | Combined[a] | 58.72 | 34 | .99 | .02 | .91 | .85–.97 |
| | Female[a] | 49.07 | 34 | .99 | .02 | .88 | .79–.97 |
| | Male | 52.42 | 34 | .97 | .04 | .98 | .87–1.11 |
| *Four factors: Carpenter et al. (1990)* | | | | | | | |
| | Combined | – | | | | | |
| | Female[a] | 54.91 | 38 | .99 | .02 | .69–.95 | .51–1.12 |
| | Male | – | | | | | |
| *Two factors: Difficulty* | | | | | | | |
| | Combined[a] | 72.77 | 53 | .99 | .02 | .82 | .75–.88 |
| | Female[a] | 68.38 | 53 | .99 | .02 | .84 | .75–.91 |
| | Male[a] | 64.25 | 53 | .99 | .02 | .77 | .64–.89 |

Blank rows indicate models could not be calculated due to the presence of Heywood cases.

[a] These models were significantly better fitting than their corresponding one-factor models. DeShon et al. (combined) $\Delta\chi^2$ (1) = 5.87, $p = .015$; DeShon et al. (female) $\Delta\chi^2$ (1) = 7.02, $p = .008$; Carpenter et al. (female) $\Delta\chi^2$ (6) = 21.00, $p = .002$; difficulty (combined) $\Delta\chi^2$ (1) = 24.76, $p < .001$; difficulty (female) $\Delta\chi^2$ (1) = 13.99, $p < .001$; and difficulty (male) $\Delta\chi^2$ (1) = 12.03, $p < .001$.

derived models, in that item classifications of 'easy' and 'hard' were determined from item threshold values, rather than derived theoretically. Therefore, the difficulty models are presented here for completeness, but will not be discussed further in this section. The interpretation of these difficulty models will be discussed in Section 3, below.

There were several appearances of Heywood cases in some multi-factor models. This is caused by negative error variance and/or a correlation between two factors exceeding 1. One cause is the estimation of too many factors and it is considered likely that this was the cause of the appearance of these cases here. Hence, the models where this occurred were considered invalid.

#### 2.2.1. Sample 1

The results from the factor analysis of Sample 1 data are presented in Table 4. In this dataset, the model based on Dillon et al. (1981) could not be calculated because there were not enough items to represent the pairwise progression factor. The results obtained for the analysis of the combined dataset show that the one-factor and DeShon et al. (1995) models had similar and acceptable fit indices. The DeShon et al. model fit statistically significantly better than the one-factor model. However, the correlation between these two factors was too high (>.90) for two separate factors to be considered meaningful. Therefore, it can be concluded that the one-factor model provided the best fit for the combined group

The comparison of the results obtained from the female and the male data give some additional insight. The male data tended to fit the one-factor model best. However, in the female group, the multi-factor models were consistently better fitting than their corresponding one-factor models, despite a high factor correlation. These results indicate that there could be a sex difference in the latent structure of this test.

#### 2.2.2. Sample 2

Table 5 presents the results from Sample 2. In calculating these models it became apparent that several pairs of items showed an empty cell in the bivariate table – that is, were statistically indistinguishable – especially in the male data. To deal with this, items from problem pairs were systematically deleted. This involved the deletion of those items that were present in the largest number of problematic pairs until there were no longer any problematic pairs left. Where this left multiple combinations, the items retained were those that resulted in the best model fit. Item 36 was not included in these models because it was almost always present in these problem pairs, and only a small number of participants successfully answered it. There were several other problem pairs, which tended to involve items 4, 9, 24, 29 and 35. No consistent characteristic of these items was found to explain this. Therefore, for Sample 2 models, the items used for male data and the female data are not directly comparable

Overall, the results indicated that, given the high correlations between factors in the multi-factor models, a one-factor model best represented the data in all groups. Unlike the Sample 1 data, the female models tended to have factors too highly correlated (.92 and .94) for the fit of the multi-factor model to be considered acceptable, although the DeShon et al. (1995) model did fit statistically significantly better than its corresponding one-factor model. However, there was a trend for the female models to have less highly correlated factors than the male models.

#### 2.2.3. Sample 3

Sample 3 results are presented in Table 6. Similar to Sample 2, there were several pairs of items in the male data set that were indistinguishable. All pairs of statistically indistinguishable items involved items 7 and 11; hence these items were excluded from the models in the male group. Again, no characteristic of these items, which they did not share with many other items that were not problematic, was found to explain this. The results show that a one-factor model provided the best fit to the data, in all groups. Although the DeShon et al. (1995) model also fit the data in the combined and female groups, the factor correlations were again high and, unlike in the other samples, this model did not fit significantly better than the corresponding one-factor model.

### 2.3. Measurement invariance and latent mean differences

#### 2.3.1. Sample 1

Although there was some evidence in this sample that the female group fit a multi-factor model, the fit of the female data to the one-factor model was also acceptable, and hence measurement invariance testing was carried out using the one-factor model. Partial measurement invariance was met by allowing the factor loading and threshold of item 3 to vary across groups while constraining equal all other

**Table 5**
Model fit indices: Sample 2.

| | Model | $\chi^2$ | *df* | CFI | RMSEA | Factor correlation | |
|---|---|---|---|---|---|---|---|
| | | | | | | Estimate | 95% CI |
| *One-Factor* | | | | | | | |
| | Combined | 755.22 | 560 | .97 | .03 | | |
| | Female | 662.90 | 560 | .98 | .02 | | |
| | Male | 462.64 | 434 | .97 | .02 | | |
| *Two factors: DeShon et al. (1995)* | | | | | | | |
| | Combined | 407.36 | 251 | .96 | .04 | .96 | .92–1.00 |
| | Female[a] | 344.68 | 251 | .96 | .03 | .94 | .87–.99 |
| | Male | 221.06 | 188 | .96 | .04 | .99 | .91–1.06 |
| *Two factors: Dillon et al. (1981)* | | | | | | | |
| | Combined | 89.07 | 76 | .99 | .02 | .95 | .86–1.02 |
| | Female | 83.54 | 76 | .99 | .02 | .92 | .82–1.01 |
| | Male | – | | | | | |
| *Two factors: Difficulty* | | | | | | | |
| | Combined[a] | 693.36 | 559 | .98 | .02 | .86 | .78–.89 |
| | Female[a] | 619.10 | 559 | .99 | .02 | .85 | .77–.90 |
| | Male[a] | 457.13 | 433 | .98 | .02 | .89 | .76–.96 |

The model based on Carpenter et al. (1990) could not be calculated due to the presence of several Heywood cases across all groups.

[a] These models were significantly better fitting than their corresponding one-factor models. DeShon et al. (female) $\Delta\chi^2$ (1) = 4.93, $p = .027$; difficulty (Combined) $\Delta\chi^2$ (1) = 34.70, $p < .001$; difficulty (female) $\Delta\chi^2$ (1) = 24.72, $p < .001$; Difficulty (Male) $\Delta\chi^2$ (1) = 5.91, $p = .015$.

parameters (see Table 7 for relevant statistics). Item 3 showed a higher loading in the male group. The latent mean in the male group was significantly higher than that of the female group (0.198, $p = .013$).

### 2.3.2. Sample 2

The models were computed excluding items 4, 9, 29, 35 and 36 due to the presence of empty cells in the bivariate table in the male data. Partial measurement invariance was met by allowing the factor loading and threshold of item 8 to vary across groups while constraining equal all other parameters (see Table 7). Item 8 showed a higher loading in the male group. The latent mean in the male group was significantly higher than that of the female group (0.270, $p = .028$).

### 2.3.3. Sample 3

The models were computed excluding items 7 and 11 due to empty cells in the bivariate table in the male data. Partial measurement invariance was met by allowing the factor loading and threshold of item 32 to vary across groups while constraining equal all other parameters (see Table 7). Item 32 showed a higher loading in the male group. The latent mean in the male group was significantly higher than that of the female group (0.295, $p = .020$).

## 2.4. Rasch analysis and differential item functioning

This section presents the results of the Rasch analysis for all samples. Fig. 1 displays the test information curves (excluding items showing poor fit; see below) for each test version. There was little difference in the test information curves for the combined, female and male groups in their respective samples. Therefore, only the curves pertaining to the combined group are displayed. As expected, the full form used in Sample 2 provided the greatest information overall.

### 2.4.1. Sample 1

One case showed misfit ($t = 6.43$) and was excluded from analysis. Examination of item fit indices resulted in the exclusion of one item (item 21 [infit =0.95, $t = -2.2$]). Given the negative t-value, this item was redundant, and therefore less of a threat to unidimensionality. There were no further misfitting items in the male or female groups when calibrated separately.

**Table 6**
Model fit indices: Sample 3.

| | Model | $\chi^2$ | *df* | CFI | RMSEA | Factor correlation | |
|---|---|---|---|---|---|---|---|
| | | | | | | Estimate | 95% CI |
| *One-factor* | | | | | | | |
| | Combined | 133.25 | 90 | .97 | .04 | | |
| | Female | 118.90 | 90 | .96 | .04 | | |
| | Male | 78.34 | 65 | .98 | .04 | | |
| *Two factors: DeShon et al. (1995)* | | | | | | | |
| | Combined | 78.56 | 53 | .98 | .04 | .92 | .81–1.02 |
| | Female | 76.54 | 53 | .96 | .04 | .93 | .78–1.06 |
| | Male | 49.22 | 34 | .97 | .06 | .95 | .81–1.08 |
| *Two factors: Dillon et al. (1981)* | | | | | | | |
| | Combined | – | | | | | |
| | Female | 19.13 | 8 | .93 | .08 | .96 | .68–1.52 |
| | Male | – | | | | | |
| *Two factors: difficulty* | | | | | | | |
| | Combined | 131.12 | 89 | .97 | .04 | .94 | .85–1.01 |
| | Female | 115.03 | 89 | .96 | .04 | .89 | .77–.98 |
| | Male[a] | 72.02 | 64 | .99 | .03 | .88 | .74–.96 |

The model based on Carpenter et al. (1990) could not be calculated due to the presence of several Heywood cases in all groups.

[a] These models were significantly better fitting than their corresponding one-factor models. Difficulty (male) $\Delta\chi^2$ (1) = 5.66, $p = .017$.

**Table 7**
Measurement invariance statistics.

| | χ2 | *df* | Δ χ2 | Δ *df* | Δ *p* | RMSEA | CFI |
|---|---|---|---|---|---|---|---|
| *Sample 1* | | | | | | | |
| Less restrictive | 164.79 | 108 | | | | .03 | .98 |
| More restrictive | 192.18 | 118 | 24.54 | 10 | .006 | .03 | .97 |
| More restrictive (partial) | 178.39 | 117 | 14.66 | 9 | .101 | .03 | .98 |
| *Sample 2* | | | | | | | |
| Less restrictive | 975.58 | 868 | | | | .02 | .98 |
| More restrictive | 1016.88 | 897 | 46.05 | 29 | .023 | .02 | .97 |
| More restrictive (partial) | 1008.28 | 896 | 37.47 | 28 | .109 | .02 | .98 |
| *Sample 3* | | | | | | | |
| Less restrictive | 165.58 | 130 | | | | .04 | .97 |
| More restrictive | 196.17 | 141 | 28.90 | 11 | .002 | .05 | .96 |
| More restrictive (partial) | 170.61 | 140 | 7.17 | 10 | .709 | .04 | .98 |

The Rasch dimension explained 36% of the variance. The first eigenvalue from the PCA of Rasch residuals was 1.09, explaining 9.9% of the residual variance. This was below the cut-off value of 1.16, supporting unidimensionality of the measure. Although the variance explained by the Rasch dimension was lower than desirable, there was little evidence of a substantial second dimension. The female and male data also showed no evidence of multidimensionality (eigenvalues of 1.11 and 1.19 respectively, with cut-off values of 1.19 and 1.31). Fig. 2 shows the ICCs for three representative items (easy, medium and difficult) in Sample 1 for the combined group.

Significant DIF was found in two items; Item 12 was easier for males, while item 18 was easier for females. The column labelled sex × item (Table 8) displays the interaction term between sex and item, and displays the female difficulty estimate, where the mid-point between female and male difficulty levels is zero. For example, the value of −0.049 for item 3 indicates that this item was easier for females, with a difference of 2 × 0.049 (0.099) logits between the difficulty for males and females.

### 2.4.2. Sample 2

Two misfitting cases were identified ($t = 10.45$ and 6.64) and excluded from analysis. Several misfitting items were identified and iteratively excluded from analysis, resulting in the exclusion of five (14%) items. Table 9 displays these items, their infit mean square and corresponding t-values, and their rule classification. Four items showed poor discrimination, while one showed redundancy (item 21). Although most of these items were classified as analytic, they represented only 33% of the total analytic items, indicating that this characteristic is unlikely to be responsible for this finding.

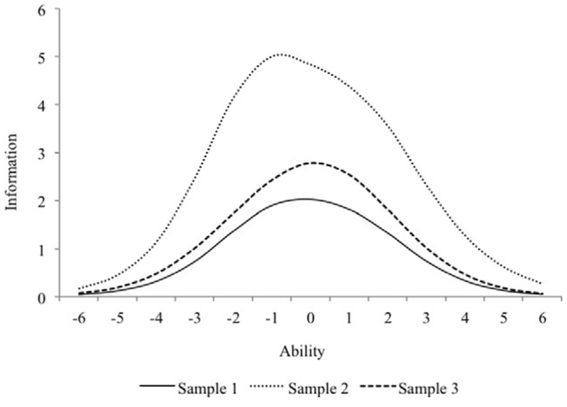


**Fig. 1.** Test information curves for the combined group for each sample.

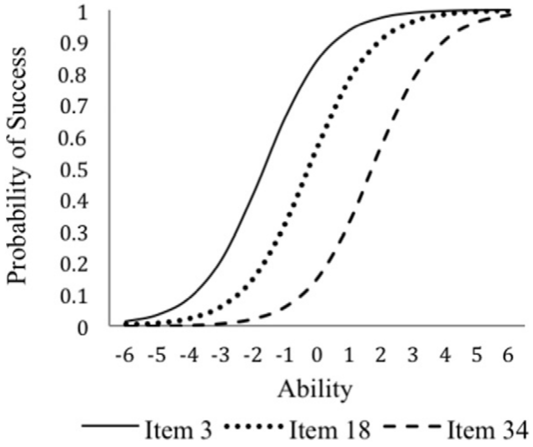


**Fig. 2.** ICCs for three representative items from Sample 1 (combined group).

The Rasch dimension explained 42% of the variance. The first eigenvalue from the PCA of Rasch residuals was 1.36, explaining 3.9% of the residual variance. This was below the cut-off value of 1.58, supporting unidimensionality of the measure. Although the variance explained by the Rasch dimension was again lower than desirable, there was little evidence of a substantial second dimension. The female and male data also showed no evidence of multidimensionality (eigenvalues of 1.45 and 1.58 respectively, with cut-off values of 1.69 and 2.20). Fig. 3 shows the ICCs for three representative items in Sample 2 for the combined group.

There were no further items showing misfit in either the male or female groups when calibrated separately. Two items showed significant DIF across sex, both of which were easier for males, with a difference of approximately .9 logits (Table 10). These items were different from those items showing DIF in Sample 1.

### 2.4.3. Sample 3

One case showed misfit ($t = 13.62$) and was excluded from analysis. All items showed acceptable fit to the Rasch model. The Rasch dimension explained 38% of the variance. The first eigenvalue from the PCA of Rasch residuals was 1.23, explaining 8.9% of the residual variance. This was below the cut-off value of 1.36, supporting unidimensionality of the measure. Although the variance explained by the Rasch dimension was lower than desirable, there was little evidence of the presence of a substantial second dimension. The female and male data also showed no evidence of multidimensionality (eigenvalues of 1.27 and 1.40 respectively, with cut-off values of 1.45 and 1.66). Fig. 4 shows the ICCs for three representative items in sample 3 for the combined group.

**Table 8**
Item fit and DIF: Sample 1.

| Item | Parameter estimate | SE | Infit | t | Sex × item | SE | Wald t |
|---|---|---|---|---|---|---|---|
| 3 | −1.665 | .084 | 1.02 | 0.3 | −0.049 | .095 | −0.52 |
| 10 | −1.266 | .076 | 0.98 | −0.4 | 0.018 | .087 | 0.21 |
| 12 | −0.824 | .069 | 0.98 | −0.6 | 0.198 | .082 | 2.41[a] |
| 15 | −1.185 | .075 | 1.03 | 0.7 | 0.038 | .086 | 0.44 |
| 16 | −0.851 | .070 | 0.97 | −0.8 | −0.073 | .078 | −0.94 |
| 18 | −0.284 | .064 | 1.01 | 0.4 | −0.160 | .071 | −2.25[a] |
| 21 | | | | | | | |
| 22 | 0.546 | .061 | 0.96 | −1.7 | 0.101 | .069 | 1.46 |
| 28 | 1.785 | .068 | 1.03 | 0.9 | 0.063 | .073 | 0.86 |
| 30 | 1.053 | .062 | 1.01 | 0.3 | −0.052 | .069 | 0.75 |
| 31 | 0.993 | .062 | 1.02 | 0.7 | −0.015 | .069 | −0.22 |
| 34 | 1.696 | .067 | 1.00 | 0.1 | −0.068 | .073 | −0.93 |

[a] Indicates the presence of significant DIF at $p < .05$.

**Table 9**
Misfitting items: Sample 2.

| Item | Infit (*95% CI*) | t | Carpenter et al. (1990) | DeShon et al. (1995) | Dillon et al. (1981) |
|---|---|---|---|---|---|
| 13 | 1.15 (*0.91–1.09*) | 3.2 | D3 | Analytic | – |
| 21 | 0.87 (*0.91–1.09*) | −3.0 | D3 | Analytic | A/S |
| 17 | 1.18 (*0.88–1.12*) | 2.7 | D3 | Analytic | PP |
| 28 | 1.14 (*0.88–1.12*) | 2.3 | D3 | Analytic | A/S |
| 20 | 1.13 (*0.89–1.11*) | 2.2 | A/S | Both | – |

*Note.* Items displayed in order of exclusion.

When the female and male data were considered separately, it was found that item 25 showed poor fit in the female group (infit = 1.13, $t = 2.2$), while in the male group no items showed poor fit. There was no significant DIF across sex in the remaining items (Table 11).

## 3. Discussion

Overall, the analyses presented largely support a unidimensional conceptualization of the APM, in the two short forms and the complete test, and in males and females. Furthermore, the analyses indicate that the APM is largely measurement invariant across sex and argue against any consistent sex differences in the item types identified to date.

Results regarding sex differences in items and item types displayed no clear pattern of association with item classifications from Carpenter et al. (1990), DeShon et al. (1995), or Dillon et al. (1981), and no clear pattern across samples. This is contrary to the findings from Mackintosh & Bennett (2005); however, it is consistent with findings from Colom and Abad (2007) and Vigneau and Bors (2008). There were significant sex differences in every item type in at least one of the samples and some individual items of all types demonstrated a significant sex difference. However, the lack of consistent results supports Vigneau and Bors' (2008) contention that the use of these taxonomies to understand sex differences in APM items may not be a valuable line of enquiry.

It was hoped that analyses of the factor structure in males and females separately would provide new insight into both the structure of the test and sex differences in performance. However, the factor analytic results were largely in line with previous research that has looked at combined data only. This research has typically reported little support for models based on Dillon et al. (1981) and DeShon et al.'s (1995) distinctions (Abad et al., 2004; Alderton & Larson, 1990; Arthur et al., 1999; Vigneau & Bors, 2008) and found the best fitting model to be a single-factor model (Abad et al., 2004; Chiesi, Ciancaleoni, Galli & Primi, 2012b. The current study suggests these findings can be extended to males and females when considered separately. There was little support for Lim's (1994) finding of a different factor structure of the APM in males and females.

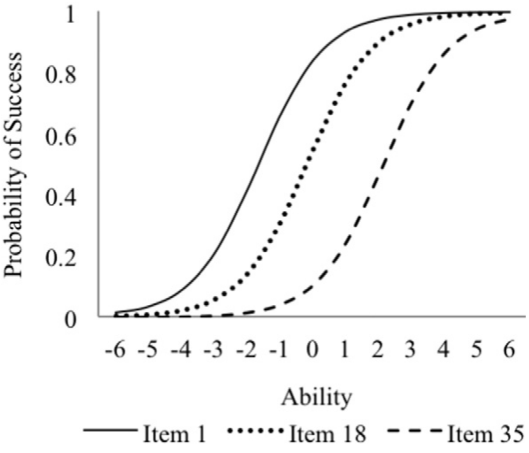


**Fig. 3.** ICCs for three representative items from Sample 2 (combined group).

While there was some evidence of a multi-factor structure in the female group, this was only found in Sample 1. It is likely that this finding occurred due to the method of item selection used to create the Bors and Stokes (1998) short form of the test. They selected those items with the highest item-total correlation, but with low inter-item correlations, in order to remove any redundancy and to obtain a wide variety of different items. This is in contrast to the short form used in Sample 3, where items with the highest item-total and inter-item correlations were selected to create a more pure measure of the APM. Therefore, the test used in Sample 1 consists of a more dissimilar group of items than the original APM, and the test used in Sample 3 consists of a more similar group of items than the original APM. These two short forms had only 40–50% of items in common, and therefore were substantially different. Consequently, the method of item selection could explain why Sample 1 showed the greatest evidence of multidimensionality and Sample 3 showed the least.

The fact that the models based on item difficulty provided arguably the best fit is difficult to interpret, but is supported by previous findings indicating that a model based on item-skewness provided the best fit (Vigneau & Bors, 2008). This result may highlight some issues regarding

**Table 10**
Item fit and DIF: Sample 2.

| Item | Parameter estimate | SE | Infit | t | Sex × item | SE | Wald t |
|---|---|---|---|---|---|---|---|
| 1 | −1.657 | .147 | 1.02 | 0.3 | −0.136 | .168 | −0.81 |
| 2 | −1.891 | .156 | 0.95 | −0.5 | 0.458 | .223 | 2.05* |
| 3 | −1.866 | .155 | 0.95 | −0.4 | −0.025 | .183 | −0.14 |
| 4 | −1.445 | .140 | 0.99 | −0.1 | 0.455 | .194 | 2.35* |
| 5 | −1.199 | .132 | 0.92 | −1.1 | −0.046 | .154 | −0.30 |
| 6 | −1.889 | .156 | 1.06 | 0.6 | 0.179 | .198 | 0.90 |
| 7 | −1.424 | .139 | 0.96 | −0.5 | 0.011 | .164 | 0.07 |
| 8 | −1.504 | .141 | 0.98 | −0.2 | −0.306 | .156 | −1.96 |
| 9 | −1.588 | .144 | 0.92 | −0.9 | 0.361 | .194 | 1.86 |
| 10 | −1.215 | .133 | 0.91 | −1.2 | 0.199 | .165 | 1.21 |
| 11 | −1.545 | .143 | 0.85 | −1.7 | −0.014 | .168 | −0.08 |
| 12 | −1.344 | .136 | 0.85 | −1.9 | 0.060 | .164 | 0.37 |
| 13 | | | | | | | |
| 14 | −1.287 | .135 | 1.02 | 0.3 | −0.207 | .151 | −1.37 |
| 15 | −0.892 | .125 | 1.01 | 0.2 | −0.275 | .138 | −1.99 |
| 16 | −0.876 | .125 | 0.94 | −0.9 | −0.144 | .141 | −1.02 |
| 17 | | | | | | | |
| 18 | −0.203 | .114 | 1.05 | 0.9 | −0.156 | .127 | −1.23 |
| 19 | −0.752 | .122 | 1.00 | 0.0 | −0.019 | .141 | −0.13 |
| 20 | | | | | | | |
| 21 | | | | | | | |
| 22 | 0.875 | .109 | 1.01 | 0.2 | 0.130 | .122 | 1.07 |
| 23 | 0.468 | .109 | 1.00 | 0.1 | −0.005 | .122 | −0.04 |
| 24 | 0.803 | .109 | 0.98 | −0.4 | 0.020 | .121 | 0.17 |
| 25 | 0.659 | .109 | 1.01 | 0.3 | 0.159 | .123 | 1.29 |
| 26 | 1.189 | .110 | 1.08 | 1.7 | 0.118 | .122 | 0.97 |
| 27 | 1.633 | .115 | 1.09 | 1.7 | −0.079 | .126 | −0.63 |
| 28 | | | | | | | |
| 29 | 2.335 | .128 | 1.12 | 1.6 | 0.096 | .136 | 0.71 |
| 30 | 1.541 | .113 | 1.01 | 0.2 | −0.176 | .125 | −1.41 |
| 31 | 1.580 | .114 | 1.00 | −0.0 | −0.116 | .126 | −0.92 |
| 32 | 1.619 | .115 | 1.02 | 0.3 | −0.056 | .126 | −0.44 |
| 33 | 1.864 | .119 | 1.07 | 1.1 | −0.243 | .131 | −1.85 |
| 34 | 1.907 | .119 | 1.01 | 0.2 | −0.248 | .132 | −1.88 |
| 35 | 2.127 | .124 | 0.91 | −1.4 | 0.086 | .132 | 0.65 |
| 36 | 4.065 | .205 | 1.03 | 0.2 | −0.081 | .220 | −0.37 |

* Indicates presence of significant DIF at *p*.

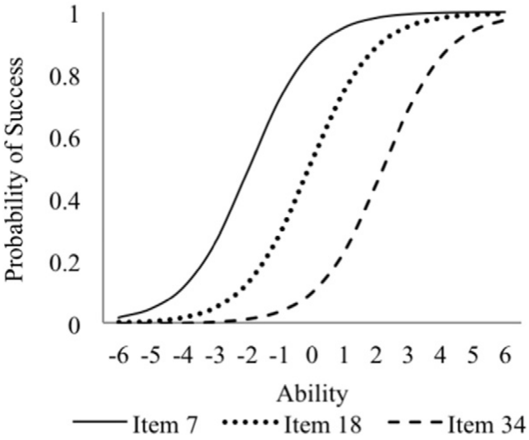


**Fig. 4.** ICCs for three representative items from Sample 3 (combined group).

the use of factor analytic methods to answer the question of dimensionality, or may be a real effect related to item position effects. There is some evidence that in addition to quantitative changes in the difficulty of items there may also be qualitative changes across the test (Vigneau & Bors, 2005), and this could also be responsible for the superior fit of the difficulty models. More research is required to disentangle the causes of any qualitative differences between the beginning and end of the test, and the influence of position effects. However, although the difficulty models fit the data well, the correlation between factors was high, particularly in Samples 2 and 3 (average correlation of .89), raising questions about the utility of considering two separate factors. Similar to the results providing some evidence of multidimensionality in the female data, the strongest evidence of multidimensionality concerning the difficulty models was found in Sample 1. Therefore, the factor analytic results indicated the presence of slight multidimensionality, but overall supported unidimensionality of the test, given the high correlations between factors in the well-fitting multi-factor models.

Interestingly, measurement invariance testing indicated that all forms were invariant across sex, with the exception of some variance in one item factor loading in each case. These loadings were higher in the male group than the female group. Therefore, there was some difference in the variance explained by the latent APM factor across sex. Additionally, latent mean difference testing supported the sex difference found in raw scores, and suggests that there is a sex difference in the latent construct measured by the APM. It is proposed that the sex difference in latent means can be interpreted one of three ways; either that males are simply better at reasoning, that visuospatial ability is involved, but was not able to be separated from reasoning ability in the current study and this is causing the difference, or that this finding is a result of the particular samples used (mainly Psychology students). Caution should be used when interpreting the results of this test until we can identify exactly why this difference occurs.

**Table 11**
Item fit and DIF: Sample 3.

| Item | Parameter estimate | SE | Infit | t | Sex × item | SE | Wald t |
|---|---|---|---|---|---|---|---|
| 7 | −1.935 | .161 | 0.97 | −0.2 | −0.167 | .173 | −0.97 |
| 11 | −2.445 | .187 | 0.91 | −0.6 | 0.333 | .234 | 1.42 |
| 13 | −0.287 | .121 | 1.08 | 1.5 | −0.146 | .132 | −1.11 |
| 15 | −1.337 | .141 | 1.01 | 0.1 | −0.023 | .155 | −0.15 |
| 16 | −0.859 | .129 | 0.98 | −0.3 | −0.011 | .142 | −0.77 |
| 17 | −0.894 | .130 | 1.02 | 0.4 | −0.164 | .140 | −1.17 |
| 18 | −0.102 | .120 | .97 | −0.6 | −0.012 | .131 | −0.09 |
| 21 | 0.431 | .119 | .96 | −0.7 | 0.094 | .129 | 0.73 |
| 23 | 0.343 | .118 | .99 | −0.2 | 0.064 | .129 | 0.50 |
| 25 | 0.122 | .119 | 1.08 | 1.7 | | | |
| 26 | 0.892 | .121 | .98 | −0.4 | −0.087 | .130 | −0.67 |
| 27 | 1.172 | .124 | .90 | −1.8 | −0.243 | .134 | −1.81 |
| 30 | 0.861 | .121 | 1.1 | 1.8 | 0.101 | .130 | 0.78 |
| 32 | 1.846 | .137 | 1.08 | 1.1 | 0.279 | .142 | 1.96 |
| 34 | 2.194 | .146 | .87 | −1.6 | −0.019 | .152 | −0.13 |

The Rasch analysis largely supported the factor analytic findings indicating unidimensionality, although again there was some evidence for slight multidimensionality. Item fit analysis indicated that the majority of items in all samples conformed to the measurement properties of the Rasch model. Between 8 and 14% of items displayed a slight deviation from the model, with 0–8% displaying high infit values, which is more problematic than low values in this case. A strict item fit criterion was used in this study, with the aim of detecting any slight departure from fit to the Rasch model. Therefore, the violations of fit were not large. The PCA of Rasch residuals also supported unidimensionality, finding no evidence of any substantial component remaining once the Rasch dimension was accounted for.

There was less evidence of significant DIF than found by Abad et al. (2004), with results more consistent with findings from Chiesi, Ciancaleoni, Galli, Morsanyi and Primi (2012a) and Chiesi, Ciancaleoni, Galli and Primi (2012b). Further, although there was some indication of significant DIF, the items identified as showing significant DIF in this study were neither the same as those showing DIF in Abad et al.'s study, nor the same as the items showing sex differences according to the chi-square tests in the current study. One interpretation of these results is that these different methods pick up on small variations between the sexes in slightly different ways. This interpretation is supported by the fact that, although there were some significant sex differences found, none of the effect sizes was particularly large and the findings were different across samples and methods. Furthermore, the fact that different items were found to show differences in the different samples can be interpreted in two ways; either the use of different forms may change the relationship between the items (as argued by Vigneau & Bors, 2005) or these analyses are picking up on slight variations within the sample, and there is no consistent difference.

Overall, it appears that, while the information processing taxonomies proposed by Carpenter, Just and Shell (1990)and DeShon, Chan and Weissbein (1995) as well as the factors proposed by Dillon et al. (1981), may be useful for understanding the ways an individual can approach solution to APM items, they may not be helpful in distinguishing between items that do or do not involve visuospatial processes, or in understanding the mechanisms behind sex differences in performance. If the test is indeed multidimensional, then it may be that the item categorisations provided so far have not identified the factors causing multidimensionality. However, given the fit of the items to the Rasch model, if multidimensionality were present, it would likely not be a large effect. The present study supports previous literature suggesting that the test is unidimensional and has expanded this finding to suggest that this unidimensionality holds in both males and females, and is invariant across sex. This indicates that the test measures the same thing in both sexes, and that an overall score is sufficient in explaining performance on all forms of the test considered in the present study.

While the results of this study support the unidimensionality of the APM, this does not mean that the APM can definitively be said to not involve visuospatial ability. Unidimensionality does not, per se, lead to the conclusion that only one ability is involved in performance. Unidimensionality simply means that all items measure the same thing. Therefore, it is entirely possible that visuospatial ability is involved in the APM and that it is involved to a similar extent in most items. Research certainly suggests that visuospatial ability could play a role in APM performance; however, its overall importance may be questionable (Schweizer et al., 2007). Given this, in order to understand better how and to what extent APM performance may be related to visuospatial ability, more research involving administration of visuospatial ability tests in conjunction with the APM is needed. Investigating the dimensionality of the APM is a useful first step, but will not answer all questions regarding the relationship between the APM, visuospatial ability and sex differences.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.paid.2015.12.008.

## References

Abad, F. J., Colom, R., Rebollo, I., & Escorial, S. (2004). Sex differential item functioning in the Raven's Advanced Progressive Matrices: Evidence for bias. *Personality and Individual Differences*, *36*, 1459–1470.

Adams, R. J., & Khoo, S. T. (1993). *Quest: The interactive test analysis system*. Melbourne, Australia: ACER.

Adams, R., Wu, M., & Wilson, M. (2012). *ACER ConQuest 3.1 [computer software]*. Melbourne, Australia: ACER.

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

Arthur, W., Tubre, T. C., Paul, D. S., & Sanchez-Ku, M. L. (1999). College-sample psychometric and normative data on a short form of the Raven Advanced Progressive Matrices test. *School Effectiveness and School Improvement*, *17*, 354–361.

Arthur, 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.

Asparouhov, T., & Muthén, B. (2006). *Robust chi square difference testing with mean and variance adjusted test statistics (Mplus Web Notes No. 10)*. Retrieved from Mplus website: http://www.statmodel.com/download/webnotes/webnote10.pdf.

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*, 382–398.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen, & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Burke, H. R. (1958). Raven's Progressive Matrices: A review and critical evaluation. *The Journal of Genetic Psychology: Child Behavior, Animal Behavior, and Comparative Psychology*, *93*, 199–228.

Burns, N. R., Bastian, V., & Nettelbeck, T. (2007). Emotional intelligence: More than personality and cognitive ability? In G. Matthews, M. Zeidner, & R. D. Roberts (Eds.), *Emotional intelligence: Knowns and unknowns*. (pp. 167–196). Oxford: Oxford University Press. Carpenter.

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*, 404–431.

Chiesi, F., Ciancaleoni, M., Galli, S., Morsanyi, K., & Primi, C. (2012a). Item response theory analysis and differential item functioning across age, gender and country of a short form of the Advanced Progressive Matrices. *Learning and Individual Differences*, *22*, 390–396.

Chiesi, F., Ciancaleoni, M., Galli, S., & Primi, C. (2012b). Using the Advanced Progressive Matrices (set I) to assess fluid ability in a short time frame: An item response theory-based analysis. *Psychological Assessment*, *24*, 892–900.

Colom, R., & Abad, F. J. (2007). Advanced Progressive Matrices and sex differences: Comment to Mackintosh and Bennett (2005). *Intelligence*, *35*, 183–185.

Colom, R., Escorial, S., & Rebollo, I. (2004). Sex differences on the Progressive Matrices are influenced by sex differences on spatial ability. *Personality and Individual Differences*, *37*, 1289–1293.

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.

Ebisch, S. J., Perrucci, M. G., Mercuri, P., Romanelli, R., Mantini, D., Romani, G. L., ... Saggino, A. (2012a). Common and unique neuro-functional basis of induction, visualization, and spatial relationships as cognitive components of fluid intelligence. *NeuroImage*, *62*, 331–342.

Gibson, W. A. (1960). Nonlinear factors in two dimensions. *Psychometrika*, *25*, 381–392.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, *9*, 139–164.

Hu, L. -T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55.

Hungi, N. (2005). Applying the Rasch model to detect biased items. In S. Alagumalai, D. D. Curtis, & N. Hungi (Eds.), *Applied Rasch measurement* (pp. 139–158). The Netherlands: Springer.

Kline, R. B. (1998). *Principles and practice of structural equation modeling*. NY: Guilford Press.

Jackson, S. A., Kleitman, S., Stankov., L., & Howie, P. (n.d.). *Individual differences in decision making depend on cognition, monitoring and control*. Unpublished manuscript.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*, 878.

Linn, M., & Petersen, A. (1985). Emergence and characterization of sex differences in spatial ability: A meta-analysis. *Child Development*, *56*, 1479–1498.

Lim, T. K. (1994). Gender-related differences in intelligence: Application of confirmatory factor analysis. *Intelligence*, *19*, 179–192.

Lynn, R., & Irwing, P. (2004a). Sex differences on the Advanced Progressive Matrices in college students. *Personality and Individual Differences*, *37*, 219–223.

Lynn, R., & Irwing, P. (2004b). Sex differences on the Progressive Matrices: A meta-analysis. *Intelligence*, *32*, 481–498.

Mackintosh, N. J., & Bennett, E. S. (2005). What do Raven's Matrices measure? An analysis in terms of sex differences. *Intelligence*, *33*, 663–674.

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence*, *37*, 1–10.

Muthén, B., Du Toit, S. H., & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes (Technical Report No. 75)*. Retrieved from Mplus website: https://www.statmodel.com/download/Article_075.pdf.

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

R Development Core Team (2014). *R [Computer software]*. Retrieved from http://www.R-project.org/.

Raîche, G. (2005). Critical eigenvalue sizes in standardized residual principal components analysis. *Rasch Measurement Transactions*, *19*, 1012.

Raven, J. (1962). *Advanced Progressive Matrices*. Oxford, England: Oxford Psychologists Press.

Schweizer, K., Goldhammer, F., Rauch, W., & Moosbrugger, H. (2007). On the validity of Raven's Matrices test: does spatial ability contribute to performance? *Personality and Individual Differences*, *43*, 1998–2010.

Vigneau, F., & Bors, D. A. (2005). Items in context: assessing the dimensionality of Raven's Advanced Progressive Matrices. *Educational and Psychological Measurement*, *65*, 109–123.

Vigneau, F., & Bors, D. A. (2008). The quest for item types based on information processing: An analysis of Raven's Advanced Progressive Matrices, with a consideration of gender differences. *Intelligence*, *36*, 702–710.

Voyer, D., Voyer, S., & Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. *Psychological Bulletin*, *117*, 250–270.

# Supplementary Materials: Paper 1

Table A1

*Factor Loadings for One-Factor Model (Sample 1)*

| Item | Combined | Female | Male |
|---|---|---|---|
| 3 | .50 | .44 | .67 |
| 10 | .60 | .61 | .57 |
| 12 | .62 | .62 | .60 |
| 15 | .49 | .53 | .37 |
| 16 | .62 | .62 | .62 |
| 18 | .56 | .56 | .60 |
| 21 | .70 | .68 | .75 |
| 22 | .67 | .64 | .73 |
| 28 | .47 | .46 | .50 |
| 30 | .56 | .54 | .61 |
| 31 | .55 | .55 | .53 |
| 34 | .53 | .56 | .44 |

*Note.* No correlated residuals were modeled.

Table A2

*Factor Loadings for DeShon et al. (1995) Model (Sample 1)*

| Item | Combined | | Female | | Male | |
|---|---|---|---|---|---|---|
| | Factor 1 | Factor 2 | Factor 1 | Factor 2 | Factor 1 | Factor 2 |
| 3 | .49 | | .43 | | .66 | |
| 10 | .62 | | .63 | | .59 | |
| 12 | .65 | | .67 | | .60 | |
| 15 | | | | | | |
| 16 | .63 | | .64 | | .62 | |
| 18 | .57 | | .57 | | .60 | |
| 21 | | .74 | | .72 | | .77 |
| 22 | .69 | | .67 | | .76 | |
| 28 | | .49 | | .48 | | .51 |
| 30 | | .57 | | .55 | | .60 |
| 31 | | | | | | |
| 34 | | .51 | | .57 | | .39 |

*Note*. No correlated residuals were modeled.

Table A3

*Factor Loadings for Carpenter et al. (1990) Model (Sample 1)*

| Item | Female | | | |
|---|---|---|---|---|
| | Factor 1 | Factor 2 | Factor 3 | Factor 4 |
| 3 | .53 | | | |
| 10 | .72 | | | |
| 12 | | | .65 | |
| 15 | | | .53 | |
| 16 | | | .63 | |
| 18 | | | | |
| 21 | | .72 | | |
| 22 | | | | .67 |
| 28 | | .48 | | |
| 30 | | | | .56 |
| 31 | | | | .57 |
| 34 | | .59 | | |

*Note.* No correlated residuals were modeled.

Table A4

*Factor Loadings for Difficulty Model (Sample 1)*

| Item | Combined | | Female | | Male | |
|---|---|---|---|---|---|---|
| | Factor 1 | Factor 2 | Factor 1 | Factor 2 | Factor 1 | Factor 2 |
| 3 | .51 | | .45 | | .69 | |
| 10 | .61 | | .62 | | .58 | |
| 12 | .63 | | .63 | | .61 | |
| 15 | .50 | | .53 | | .37 | |
| 16 | .63 | | .63 | | .63 | |
| 18 | .57 | | .57 | | .61 | |
| 21 | .71 | | .69 | | .77 | |
| 22 | .68 | | .65 | | .75 | |
| 28 | | .52 | | .50 | | .57 |
| 30 | | .62 | | .59 | | .71 |
| 31 | | .61 | | .60 | | .61 |
| 34 | | .58 | | .61 | | .51 |

*Note.* No correlated residuals were modeled.

Table A5

*Factor Loadings for One-Factor Model (Sample 2)*

| Item | Combined | Female | Male |
|---|---|---|---|
| 1 | .61 | .63 | .58 |
| 2 | .69 | .72 | .58 |
| 3 | .64 | .60 | .74 |
| 4 | .67 | .70 | |
| 5 | .70 | .75 | .65 |
| 6 | .60 | .60 | .60 |
| 7 | .68 | .68 | .69 |
| 8 | .66 | .61 | .85 |
| 9 | .71 | .69 | |
| 10 | .76 | .78 | .71 |
| 11 | .78 | .78 | .78 |
| 12 | .79 | .78 | .84 |
| 13 | .48 | .50 | .40 |
| 14 | .62 | .61 | .70 |
| 15 | .59 | .62 | .57 |
| 16 | .71 | .70 | .77 |
| 17 | .42 | .36 | .59 |
| 18 | .60 | .54 | .75 |
| 19 | .63 | .70 | .36 |
| 20 | .51 | .52 | .45 |
| 21 | .76 | .76 | .75 |
| 22 | .62 | .63 | .57 |
| 23 | .63 | .56 | .82 |
| 24 | .67 | .64 | .70 |
| 25 | .61 | .62 | .54 |
| 26 | .53 | .55 | .43 |
| 27 | .55 | .54 | .60 |
| 28 | .41 | .39 | .45 |
| 29 | .45 | .43 | |
| 30 | .58 | .59 | .57 |
| 31 | .60 | .59 | .65 |
| 32 | .57 | .56 | .60 |
| 33 | .50 | .53 | .42 |
| 34 | .57 | .62 | .40 |
| 35 | .72 | .69 | |

*Note.* No correlated residuals were modeled.

Table A6

*Factor Loadings for DeShon et al. (1995) Model (Sample 2)*

| Item | Combined | | Female | | Male | |
|---|---|---|---|---|---|---|
| | Factor 1 | Factor 2 | Factor 1 | Factor 2 | Factor 1 | Factor 2 |
| 1 | | .61 | | .64 | | .58 |
| 2 | | | | | | |
| 3 | .67 | | .64 | | .74 | |
| 4 | | .65 | | .70 | | |
| 5 | | | | | | |
| 6 | | | | | | |
| 7 | .69 | | .70 | | .67 | |
| 8 | | .68 | | .62 | | .87 |
| 9 | .72 | | .70 | | | |
| 10 | .76 | | .78 | | .73 | |
| 11 | .79 | | .78 | | .79 | |
| 12 | .80 | | .79 | | .84 | |
| 13 | | .49 | | .52 | | .40 |
| 14 | | | | | | |
| 15 | | | | | | |
| 16 | .71 | | .70 | | .77 | |
| 17 | | .43 | | .38 | | .60 |
| 18 | .61 | | .57 | | .73 | |
| 19 | | | | | | |
| 20 | | | | | | |
| 21 | | .77 | | .77 | | .74 |
| 22 | .63 | | .63 | | .57 | |
| 23 | .64 | | .57 | | .83 | |
| 24 | .66 | | .64 | | .71 | |
| 25 | | | | | | |
| 26 | | | | | | |
| 27 | | .56 | | .55 | | .60 |
| 28 | | .41 | | .41 | | .45 |
| 29 | | .45 | | .44 | | |
| 30 | | .59 | | .61 | | .57 |
| 31 | | | | | | |
| 32 | .56 | | .55 | | .60 | |
| 33 | .49 | | .54 | | .39 | |
| 34 | | .57 | | .64 | | .40 |
| 35 | | | | | | |
| 36 | | | | | | |

*Note.* No correlated residuals were modeled.

Table A7

*Factor Loadings for Dillon et al. (1981) Model (Sample 2)*

| Item | Combined | | Female | |
|---|---|---|---|---|
| | Factor 1 | Factor 2 | Factor 1 | Factor 2 |
| 1 | | | | |
| 2 | .74 | | .77 | |
| 3 | .66 | | .62 | |
| 4 | .72 | | .75 | |
| 5 | .73 | | .76 | |
| 6 | | | | |
| 7 | | .73 | | .74 |
| 8 | | | | |
| 9 | | .72 | | .69 |
| 10 | | .77 | | .79 |
| 11 | | .79 | | .78 |
| 12 | | | | |
| 13 | | | | |
| 14 | | | | |
| 15 | | | | |
| 16 | | .69 | | .70 |
| 17 | .44 | | .42 | |
| 18 | | | | |
| 19 | | | | |
| 20 | | | | |
| 21 | | .76 | | .76 |
| 22 | | | | |
| 23 | | | | |
| 24 | | | | |
| 25 | | | | |
| 26 | .54 | | .56 | |
| 27 | | | | |
| 28 | | .37 | | .38 |
| 29 | | | | |
| 30 | | | | |
| 31 | | | | |
| 32 | | | | |
| 33 | | | | |
| 34 | | | | |
| 35 | | .68 | | .65 |
| 36 | | | | |

*Note.* No correlated residuals were modeled.

Table A8

*Factor Loadings for Difficulty Model (Sample 2)*

| Item | Combined | | Female | | Male | |
|---|---|---|---|---|---|---|
| | Factor 1 | Factor 2 | Factor 1 | Factor 2 | Factor 1 | Factor 2 |
| 1 | .62 | | .64 | | .58 | |
| 2 | .70 | | .73 | | .58 | |
| 3 | .65 | | .61 | | .74 | |
| 4 | .68 | | .72 | | | |
| 5 | .71 | | .76 | | .56 | |
| 6 | .61 | | .61 | | .61 | |
| 7 | .69 | | .69 | | .70 | |
| 8 | .67 | | .62 | | .86 | |
| 9 | .72 | | .70 | | | |
| 10 | .77 | | .78 | | .72 | |
| 11 | .79 | | .79 | | .79 | |
| 12 | .80 | | .79 | | .85 | |
| 13 | .49 | | .51 | | .41 | |
| 14 | .63 | | .62 | | .71 | |
| 15 | .60 | | .63 | | .57 | |
| 16 | .72 | | .71 | | .77 | |
| 17 | .42 | | .37 | | .59 | |
| 18 | .61 | | .56 | | .76 | |
| 19 | .64 | | .71 | | .37 | |
| 20 | .51 | | .53 | | .45 | |
| 21 | .78 | | .77 | | .76 | |
| 22 | | .66 | | .67 | | .60 |
| 23 | .64 | | .57 | | .83 | |
| 24 | | .71 | | .68 | | .74 |
| 25 | | .65 | | .67 | | .58 |
| 26 | | .57 | | .59 | | .45 |
| 27 | | .58 | | .57 | | .63 |
| 28 | | .43 | | .42 | | .48 |
| 29 | | .48 | | .47 | | |
| 30 | | .62 | | .63 | | .61 |
| 31 | | .65 | | .63 | | .69 |
| 32 | | .61 | | .59 | | .63 |
| 33 | | .54 | | .57 | | .45 |
| 34 | | .61 | | .67 | | .43 |
| 35 | | .76 | | .73 | | |
| 36 | | | | | | |

*Note.* No correlated residuals were modeled.

Table A9

*Factor Loadings for One-Factor Model (Sample 3)*

| Item | Combined | Female | Male |
|---|---|---|---|
| 7 | .61 | .55 | |
| 11 | .65 | .62 | |
| 13 | .46 | .51 | .40 |
| 15 | .56 | .65 | .40 |
| 16 | .62 | .67 | .54 |
| 17 | .53 | .60 | .47 |
| 18 | .67 | .58 | .83 |
| 21 | .69 | .66 | .72 |
| 23 | .63 | .53 | .81 |
| 25 | .50 | .36 | .77 |
| 26 | .64 | .60 | .71 |
| 27 | .73 | .69 | .83 |
| 30 | .47 | .44 | .50 |
| 32 | .54 | .32 | .80 |
| 34 | .82 | .78 | .88 |

*Note.* No correlated residuals were modeled.

Table A10

*Factor Loadings for DeShon et al. (1995) Model (Sample 3)*

| Item | Combined | | Female | | Male | |
|---|---|---|---|---|---|---|
| | Factor 1 | Factor 2 | Factor 1 | Factor 2 | Factor 1 | Factor 2 |
| 7 | .64 | | .60 | | | |
| 11 | .66 | | .63 | | | |
| 13 | | .43 | | .49 | | .36 |
| 15 | | | | | | |
| 16 | .63 | | .68 | | .55 | |
| 17 | | .52 | | .56 | | .49 |
| 18 | .70 | | .62 | | .82 | |
| 21 | | .67 | | .62 | | .74 |
| 23 | .66 | | .57 | | .83 | |
| 25 | | | | | | |
| 26 | | | | | | |
| 27 | | .76 | | .72 | | .85 |
| 30 | | .50 | | .46 | | .56 |
| 32 | .58 | | .39 | | .82 | |
| 34 | | .83 | | .80 | | .90 |

*Note.* No correlated residuals were modeled.

Table A11

*Factor Loadings for Dillon et al. (1981) Model (Sample 3)*

| | Female | |
|---|---|---|
| | Factor 1 | Factor 2 |
| 7 | | .52 |
| 11 | | .65 |
| 13 | | |
| 15 | | |
| 16 | | .72 |
| 17 | .58 | |
| 18 | | |
| 21 | | .71 |
| 23 | | |
| 25 | | |
| 26 | .57 | |
| 27 | | |
| 30 | | |
| 32 | | |
| 34 | | |

*Note.* No correlated residuals were modeled.

Table A12

*Factor Loadings for Difficulty Model (Sample 3)*

| | Combined | | Female | | Male | |
|---|---|---|---|---|---|---|
| | Factor 1 | Factor 2 | Factor 1 | Factor 2 | Factor 1 | Factor 2 |
| 7 | .62 | | .56 | | | |
| 11 | .65 | | .64 | | | |
| 13 | .47 | | .52 | | .41 | |
| 15 | .57 | | .67 | | .41 | |
| 16 | .63 | | .69 | | .56 | |
| 17 | .54 | | .62 | | .59 | |
| 18 | .68 | | .59 | | .86 | |
| 21 | | .70 | | .68 | .75 | |
| 23 | .64 | | | .54 | .83 | |
| 25 | .51 | | .37 | | .80 | |
| 26 | | .65 | | .61 | | .74 |
| 27 | | .74 | | .71 | | .86 |
| 30 | | .48 | | .45 | | .52 |
| 32 | | .55 | | .33 | | .83 |
| 34 | | .83 | | .80 | | .91 |

*Note.* No correlated residuals were modeled.

# Appendix B

## Supplementary Materials: Paper 2

### Incremental validity analysis of manifest scores

Analyses of incremental validity were performed in R (R Core Team, 2015) using the base stats package. Regression models including the alternate Gf measure only and the alternate Gf measure as well as the Gv measures were calculated. ANOVA was used to test for incremental validity of the Gv measures. The Gv variable included in Model 3 from Sample 2 and 3 was calculated by summing the z-score of the individual Gv measures.

### Sample 1

Table B1

*Regression models for Sample 1*

| | B | SE B | *β* | *t* |
|---|---|---|---|---|
| Model 1 | | | | |
| EA | .70 | .06 | .54*** | 12.09 |
| Model 2 | | | | |
| EA | .41 | .06 | .32*** | 7.30 |
| Rotation | .50 | .04 | .48*** | 11.03 |

*Note*. Model 1 $R^2$ = .29, *F*(1, 351) = 146.1, p < .001; Model 2 $R^2$ = .48 F (2,350) = 159, p < .001. EA = Esoteric Analogies.

The ANOVA demonstrated that the 19% increase in explained variance in Model 2 was statistically significant *F*(1,351) = 121.74, *p* < .001.

**Sample 2**

Table B2

*Regression models for Sample 2*

| | B | SE B | *β* | *t* |
|---|---|---|---|---|
| Model 1 | | | | |
| CAB-I | .57 | .07 | .47*** | 8.08 |
| Model 2 | | | | |
| CAB-I | .42 | .07 | .34*** | 5.61 |
| PF | .08 | .03 | .17* | 2.37 |
| MRT | .11 | .04 | .21** | 3.05 |
| Model 3 | | | | |
| CAB-I | .41 | .07 | .33*** | 5.64 |
| Gv | .45 | .08 | .33*** | 5.60 |

*Note.* Model 1 $R^2$ = .22, *F*(1, 234) = 65.24, p < .001; Model 2 $R^2$ = .31 *F*(3,232) = 34.92, *p* < .001; Model 3 $R^2$ = .31 *F*(2,233) = 52.54, *p* < .001. CAB-I = Comprehensive Ability Battery – Inductive Reasoning; MRT = Mental Rotations Test; PF = Space Relations: Paper Folding

The ANOVAs demonstrated that the 9% increase in explained variance in Models 2 and 3 was statistically significant (Model 2: *F*[2,232] = 15.67, p < .001; Model 3: *F*[1,233] = 31.37, p < .001).

**Sample 3**

Table B3

*Regression models for Sample 3 (SPM)*

| | B | SE B | *β* | *t* |
|---|---|---|---|---|
| Model 1 | | | | |
| CAB-I | 2.81 | .20 | .64*** | 13.95 |
| Model 2 | | | | |
| CAB-I | 1.85 | .21 | .42*** | 8.75 |
| PF | .28 | .12 | .11* | 2.33 |
| MRT | .49 | .18 | .13** | 2.70 |
| CAB-Cf | .84 | .15 | .28*** | 5.62 |
| Model 3 | | | | |
| CAB-I | 1.91 | .21 | .43*** | 9.06 |
| Gv | 1.52 | .18 | .40*** | 8.32 |

*Note.* Model 1 $R^2$ = .41, *F*(1, 285) = 194.5, p < .001; Model 2 $R^2$ = .53 *F*(4, 282) = 80.3, p < .001; Model 3 $R^2$ = .52 *F*(2,284) = 155.2, p < .001. CAB-I = Comprehensive Ability Battery – Inductive Reasoning; CAB-Cf = Comprehensive Ability Battery – Flexibility of Closure; MRT = Mental Rotations Test; PF = Space Relations: Paper Folding

The ANOVAs demonstrated that the 12% increase in explained variance in Model 2 was statistically significant *F*(3, 282) = 25.50, p < .001, as was the 11% increase in variance explained in Model 3, *F*(1,284) = 69.24, p < .001.

## References

R Core Team (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

# Appendix C

## Supplementary Materials: Paper 3

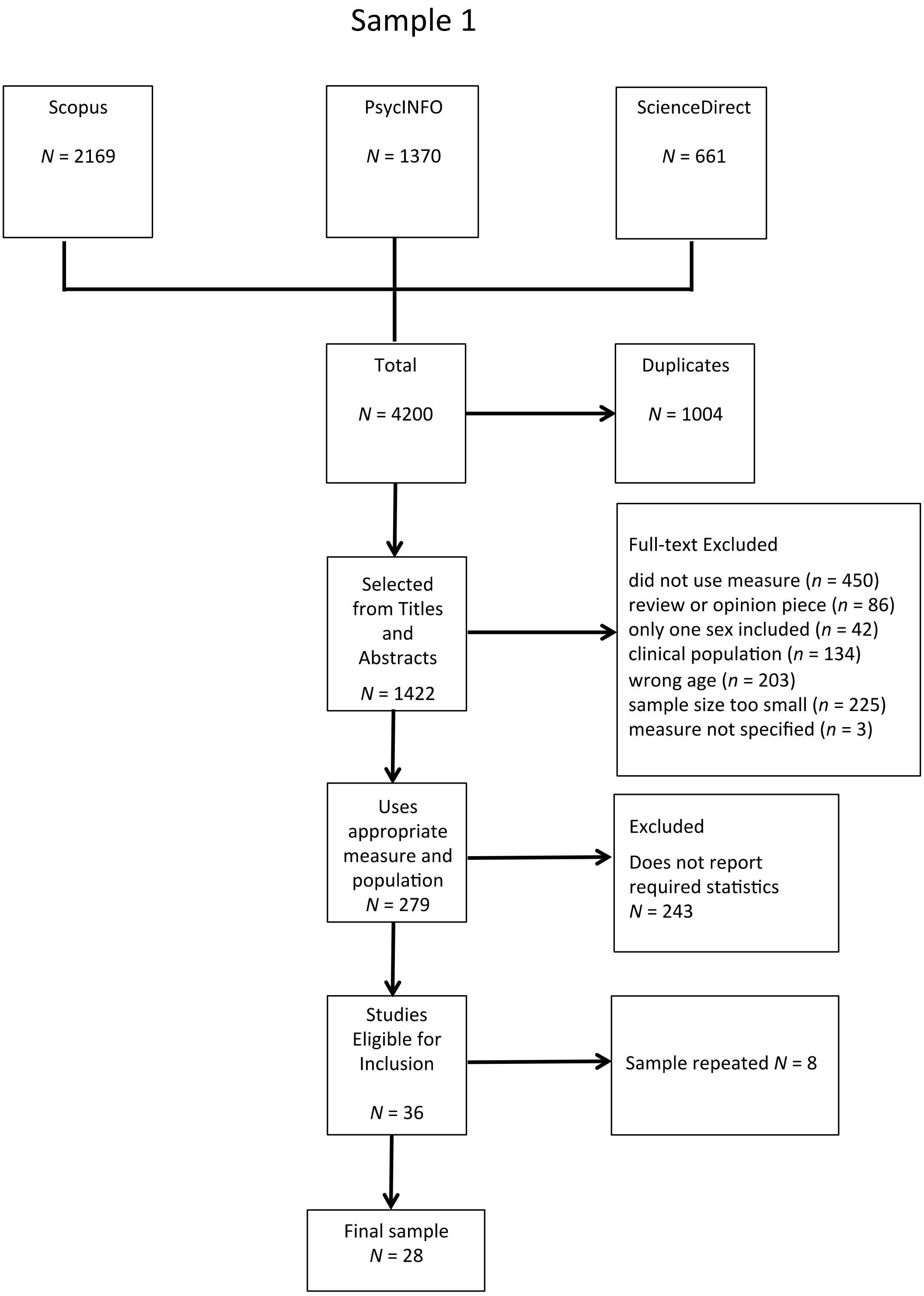

Sample 1
Scopus
*N* = 2169
PsycINFO
*N* = 1370
ScienceDirect
*N* = 661
Total
*N* = 4200
Duplicates
*N* = 1004
Selected from Titles and Abstracts
*N* = 1422
Full-text Excluded
did not use measure (*n* = 450)
review or opinion piece (*n* = 86)
only one sex included (*n* = 42)
clinical population (*n* = 134)
wrong age (*n* = 203)
sample size too small (*n* = 225)
measure not specified (*n* = 3)
Uses appropriate measure and population
*N* = 279
Excluded
Does not report required statistics
*N* = 243
Studies Eligible for Inclusion
*N* = 36
Sample repeated *N* = 8
Final sample
*N* = 28

*Figure C1*. PRISMA diagram for Sample 1 search.

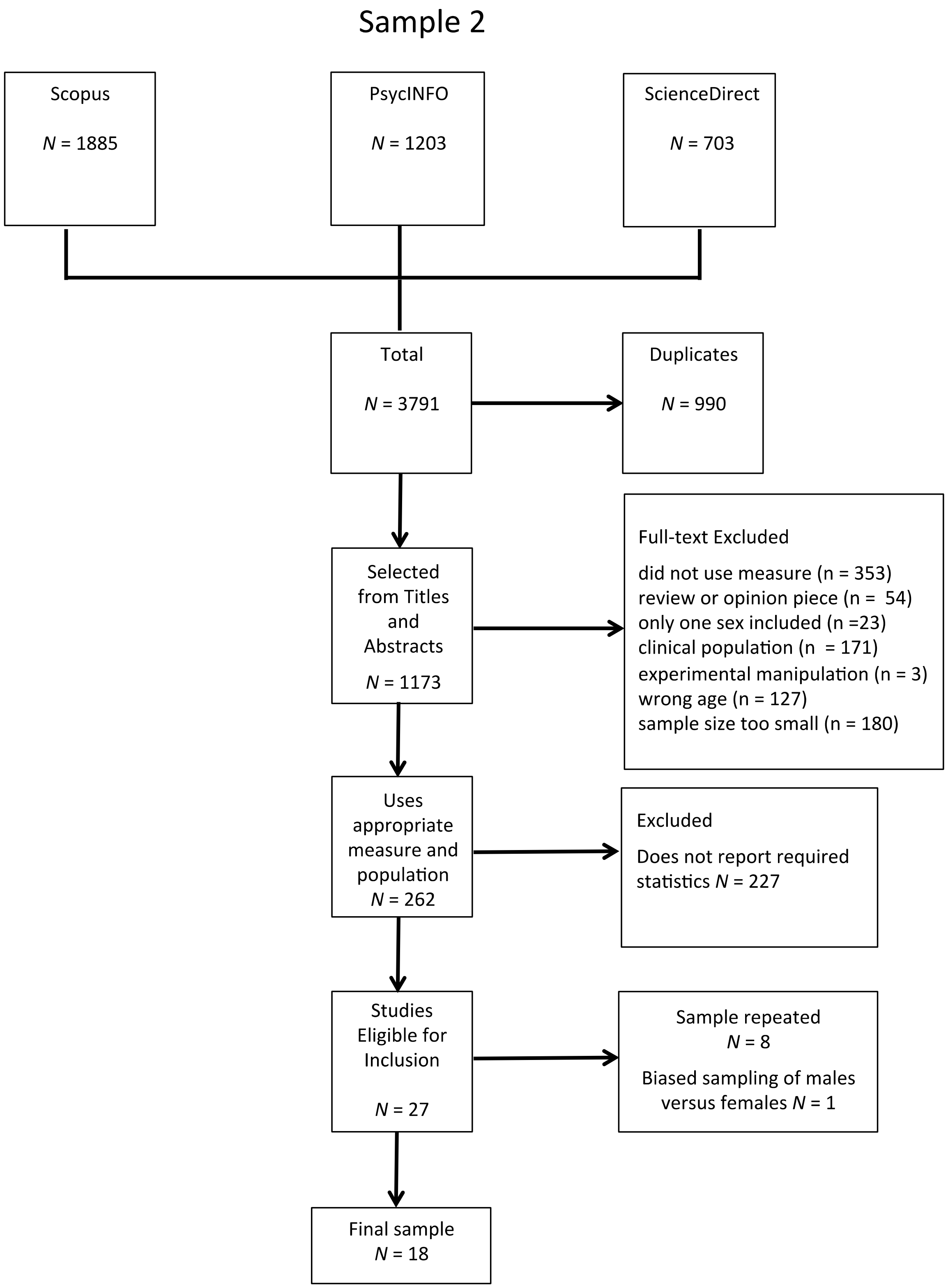

Sample 2
Scopus
N = 1885
PsycINFO
N = 1203
ScienceDirect
N = 703
Total
N = 3791
Duplicates
N = 990
Selected from Titles and Abstracts
N = 1173
Full-text Excluded
did not use measure (n = 353)
review or opinion piece (n = 54)
only one sex included (n =23)
clinical population (n = 171)
experimental manipulation (n = 3)
wrong age (n = 127)
sample size too small (n = 180)
Uses appropriate measure and population
N = 262
Excluded
Does not report required statistics N = 227
Studies Eligible for Inclusion
N = 27
Sample repeated
N = 8
Biased sampling of males versus females N = 1
Final sample
N = 18

*Figure C2*. PRISMA diagram for Sample 2 search.

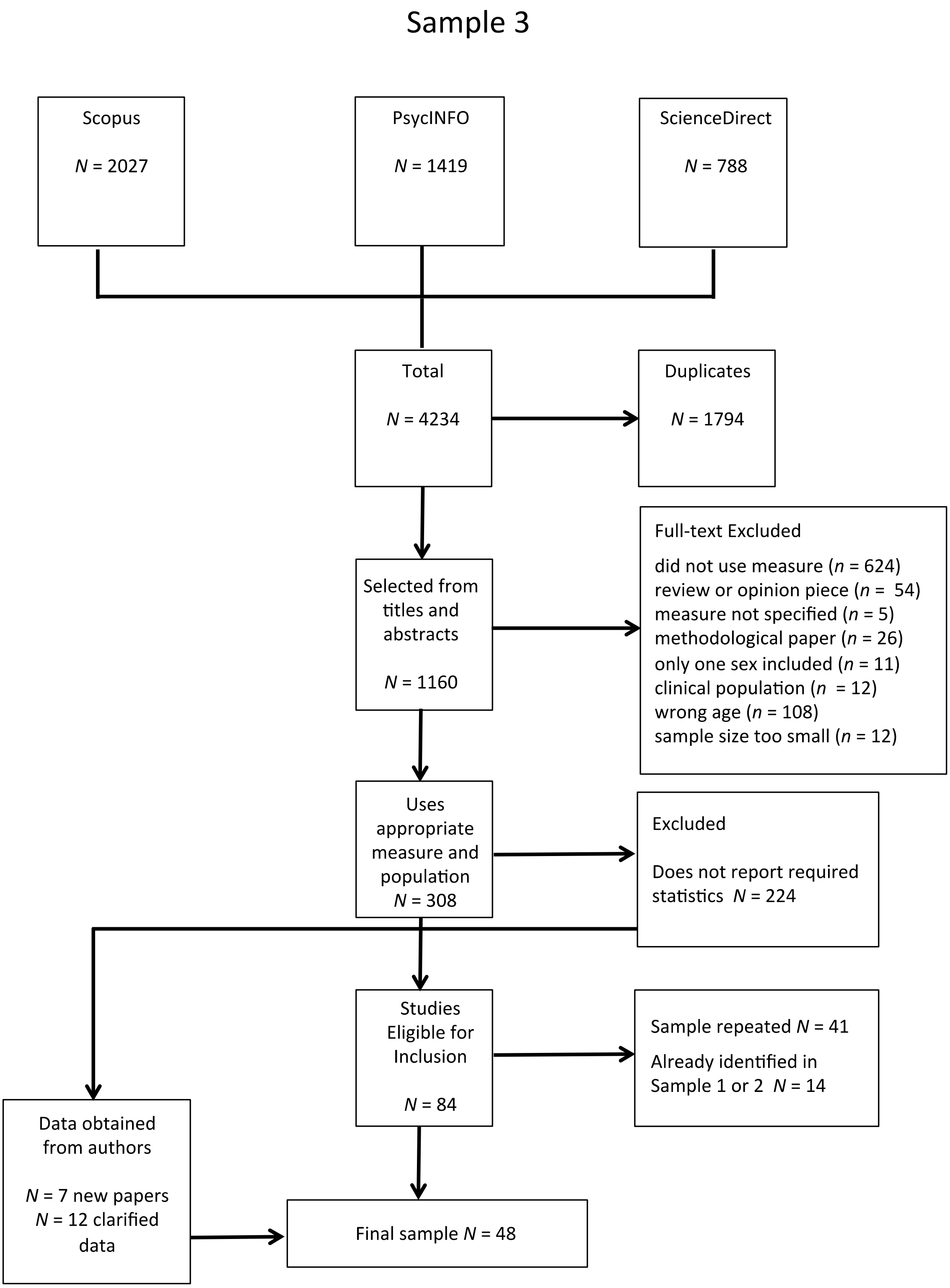

Sample 3
Scopus
N = 2027
PsycINFO
N = 1419
ScienceDirect
N = 788
Total
N = 4234
Duplicates
N = 1794
Selected from titles and abstracts
N = 1160
Full-text Excluded
did not use measure (n = 624)
review or opinion piece (n = 54)
measure not specified (n = 5)
methodological paper (n = 26)
only one sex included (n = 11)
clinical population (n = 12)
wrong age (n = 108)
sample size too small (n = 12)
Uses appropriate measure and population
N = 308
Excluded
Does not report required statistics N = 224
Studies Eligible for Inclusion
N = 84
Sample repeated N = 41
Already identified in Sample 1 or 2 N = 14
Data obtained from authors
N = 7 new papers
N = 12 clarified data
Final sample N = 48

*Figure C3.* PRISMA diagram for Sample 3 search.

# Appendix D

## Supplementary Materials: Paper 4

Table D1

*Factor loadings and communalities for the Abstract Reasoning and Numerical Reasoning subtests*

| Item | Loading | $h^2$ | Item | Loading | $h^2$ |
|---|---|---|---|---|---|
| AR1 | **.50** | .25 | NR1 | **.57** | .33 |
| AR2 | **.68** | .47 | NR2 | **.82** | .67 |
| AR3 | .25 | .06 | NR3 | **.65** | .43 |
| AR4 | **.38** | .14 | NR4 | **.71** | .50 |
| AR5 | **.44** | .20 | NR5 | **.64** | .41 |
| AR6 | .29 | .09 | NR6 | **.51** | .26 |
| AR7 | **.73** | .54 | NR7 | **.65** | .42 |
| AR8 | **.50** | .26 | NR8 | **.66** | .44 |
| AR9 | **.55** | .30 | NR9 | **.51** | .26 |
| AR10 | **.80** | .65 | NR10 | **.58** | .33 |
| AR11 | **.62** | .38 | NR11 | **.69** | .47 |
| AR12 | **.65** | .43 | NR12 | **.73** | .54 |
| AR13 | **.61** | .38 | NR13 | **.63** | .40 |
| AR14 | **.33** | .11 | NR14 | **.66** | .43 |
| AR15 | **.57** | .33 | NR15 | **.65** | .42 |
| AR16 | **.74** | .56 | NR16 | **.78** | .61 |
| AR17 | **.66** | .44 | NR17 | **.72** | .51 |
| AR18 | **.52** | .28 | NR18 | **.75** | .57 |
| AR19 | **.73** | .54 | NR19 | **.63** | .39 |
| AR20 | **.64** | .42 | NR20 | **.74** | .55 |
| AR21 | **.58** | .34 | NR21 | **.81** | .65 |
| AR22 | **.59** | .36 | NR22 | **.87** | .75 |
| AR23 | **.52** | .27 | NR23 | **.75** | .56 |
| AR24 | **.53** | .29 | NR24 | **.77** | .59 |
| AR25 | **.61** | .38 | NR25 | **.60** | .36 |

*Note.* Factor loadings >.30 in bold.

Table D2

*Factor loadings and communalities for the Verbal Reasoning subtest*

| | VR Scale 3-Factors | | | | VR Scale 2-Factors | | |
|---|---|---|---|---|---|---|---|
| Item | I | II | III | $h^2$ | I | II | $h^2$ |
| VR1 | **.62** | **.40** | **-.33** | .44 | **.67** | **-.30** | .31 |
| VR2 | | | .23 | .15 | | | |
| VR3 | .24 | | | .11 | | | |
| VR4 | .26 | **.31** | | .23 | **.31** | | .16 |
| VR5 | **.48** | **.31** | | .37 | **.52** | | .30 |
| VR6 | **.45** | .29 | | .31 | **.49** | | .25 |
| VR7 | **.71** | | | .40 | **.71** | | .39 |
| VR8 | | .26 | | .12 | | | |
| VR9 | **.35** | | | .20 | **.36** | | .17 |
| VR10 | .28 | | .25 | .21 | .27 | .25 | .21 |
| VR11 | **.68** | | | .40 | **.68** | | .40 |
| VR12 | **.63** | **.31** | | .53 | **.68** | | .48 |
| VR13 | | | .26 | .15 | | .25 | .12 |
| VR14 | .23 | | **.35** | .26 | .22 | **.35** | .26 |
| VR15 | | | **.37** | .24 | | **.38** | .23 |
| VR16 | **.56** | | | .33 | **.54** | | .31 |
| VR17 | | **.40** | **.53** | .49 | | **.54** | .32 |
| VR18 | **.38** | | | .17 | **.39** | | .17 |
| VR19 | | | | .02 | | | |
| VR20 | **.39** | | | .14 | **.40** | | .13 |
| VR21 | **.30** | .22 | .28 | .35 | **.33** | .29 | .31 |
| VR22 | **.64** | | | .46 | **.61** | | .41 |
| VR23 | **.55** | | .20 | .47 | **.55** | .20 | .46 |
| VR24 | | | **.52** | .29 | | **.51** | .26 |
| VR25 | **.47** | | | .21 | **.48** | | .20 |
| VR26 | .26 | -.24 | .22 | .21 | .23 | .21 | .15 |
| VR27 | **.35** | | | .16 | **.35** | | .16 |
| VR28 | | | | .10 | | | |
| VR29 | | | **.36** | .22 | | **.38** | .22 |
| VR30 | | | **.51** | .41 | | **.53** | .40 |
| VR31 | | | **.79** | .64 | | **.80** | .64 |
| VR32 | **.54** | | | .29 | **.51** | | .26 |
| VR33 | .23 | | **.51** | .44 | .21 | **.52** | .44 |
| VR34 | | | **.60** | .35 | | **.62** | .34 |
| VR35 | **.30** | | **.36** | .34 | **.30** | **.36** | .34 |

Factor Correlations

| | I | II | | | I |
|---|---|---|---|---|---|
| II | .09 | | | II | 0.58 |
| III | .56 | .14 | | | |

*Note.* Factor loadings < .20 not displayed. Factor loadings >.30 in bold.

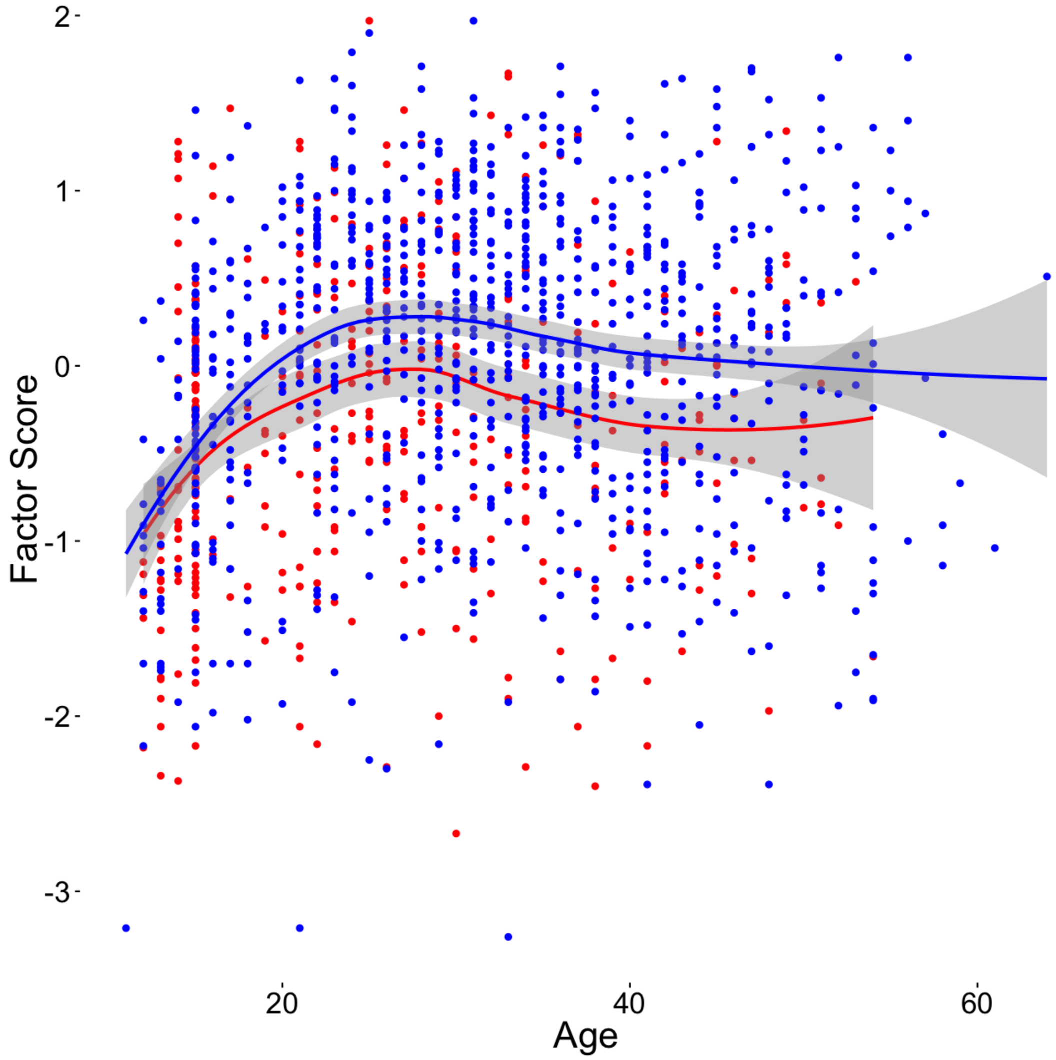


*Figure D1*. Factor scores on Factor I by age and sex (females = red, males = blue). Lines indicate the smoothed conditional mean for each factor, with confidence bands indicating the 95% confidence interval.

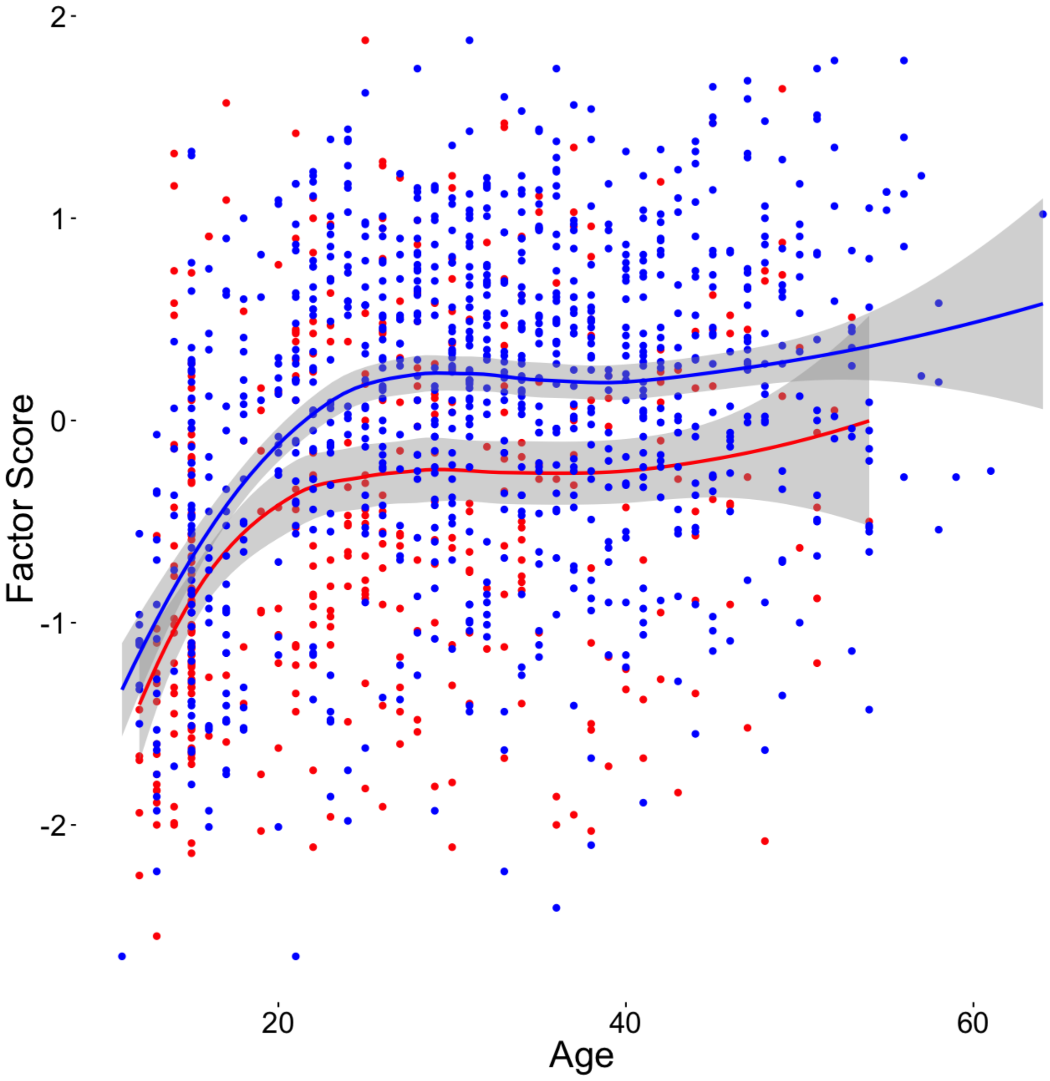


*Figure D2.* Factor scores on Factor II by age and sex (females = red, males = blue). Lines indicate the smoothed conditional mean for each factor, with confidence bands indicating the 95% confidence interval.