Sex Differences in General Intelligence

A Psychometric Investigation of Group Differences in Mean and Variability as Measured by The Raven's Standard Progressive Matrices

Emily Savage-M^{c}Glynn

This Dissertation is submitted for the degree of Doctor of Philosophy at the University of Cambridge

December 2010

PREFACE

This dissertation is my own work and contains nothing which is the outcome of word done in collaboration with others, except as specified in the text and Acknowledgements.

This dissertation does not exceed 80,000 words.

Sex Differences in General Intelligence:

A Psychometric Investigation of Group Differences in Mean and Variability as Measured by the Raven's Standard Progressive Matrices

Emily Savage-M^{c}Glynn, Emmanuel College

Researchers and the general public alike continue to debate 'which is the smarter sex?' Research to date suggests that males outperform females, females outperform males, while others find no differences in mean or variance. These inconsistent results are thought to occur for two reasons. First, studies rely on opportunity samples rather than samples that represent the general population. Second, researchers have not availed themselves of advances in psychometrics that allow for identification of bias in test items and the reliable evaluation of group differences. This dissertation addresses these two identified needs in the literature.

Using a large representative U.K. sample, 926 seven to 18 year olds were assessed with the Raven's Standard Progressive Matrices Plus (SPM+), a measure considered to be one of the best measures of general intelligence. In assessing a one-factor model of general intelligence, four research aims were addressed. First, confirmatory factor analyses and assessment of measurement invariance revealed that the SPM+ is not biased to either sex. Second, multiple group confirmatory factor analyses revealed there to be no significant differences between males and females in either mean or variance. Third, analyses revealed no significant sex differences in mean or variability in younger or older participants. Finally, method effects of Gestalt and Visuospatial answering strategies explained some of the residual variance in the model. For the overall sample, males were significantly disadvantaged by the visuospatial element of some of the items. For older participants, the influence of the methods effects was equivalent.

It can generally be concluded that there are no significant sex differences in mean or variability on the SPM+ suggesting that there is no sex difference in general intelligence. Future research should employ representative samples and robust statistical methodologies to assess sex differences on the Raven's from a multiple factor perspective.

Acknowledgements

There are many people that I wish to thank for the help and support that I have received throughout the last four years, both professionally and personally.

First, this research was supported by the Cambridge Assessment Doctoral Studentship in Psychometrics. Further funding support has been gratefully provided by Emmanuel College. I would also like to extend immense thanks to the support of Pearson Assessment for the use of the Raven's SPM data.

I extend great thanks to all of the study participants, their families and schools for taking part in the standardisation project. Further thanks to Lyn Dale and Vikas Dhawan who were administrators on the project and to the test examiners who collected data across the U.K. Special thanks to Dr. Paul Crane and Dr. David Good with whom I consulted in the early phases of this project and, with their guidance, feel that my work has been strengthened. My gratitude and appreciation to Irenee Daly, who has helped to make the process of doing a Ph.D. much more enjoyable, and in our joint attempts to become better writers, has read and improved many of these chapters.

My thanks and appreciation goes to Prof. Melissa Hines and Prof. John Rust who have supervised and guided me through my doctoral studies. My sincere appreciation and gratitude goes to Dr. Rosie Ensor for her thoughtfulness, insight and patience, without whom I'd still be agonising over Mplus and my analysis.

I extend my heartfelt thanks and gratitude to Lyn Dean for looking after Brendan for the last 3 years. Knowing he was so well cared for ensured I could focus on my studies, and ultimately get this PhD finished!

To my wonderful family: I thank you for your constant love, support

and understanding through this epic adventure. I couldn't have done it without you.

Finally, I must thank two very special people to whom I dedicate this dissertation. Kevin, you've been with me through every step of this very, very long journey. Words cannot express how thankful I am and how much I appreciate your support of my dreams – even when I couldn't remember what they were or why I was striving for them. You constantly reminded me of the light at the end of the tunnel, even when I couldn't see it.

To my Brendan Moo – in your own way, you have helped me through this Ph.D. more than anyone with your endless hugs, kisses and smiles. You make every day worthwhile, and in completing this, I hope to have shown you that any dream is possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

List of Equations

Glossary of Acronyms

Acronym	*Definition*
1PL	One Parameter Logistic Model
2PL	Two Parameter Logistic Model
APM	Advanced Progressive Matrices
CAT	Cognitive Abilities Test
CFA	Confirmatory Factor Analysis
CFI	Comparative Fit Index
CPM	Coloured Progressive Matrices
CTT	Classical Test Theory
DAT	Differential Aptitude Test
DIF	Differential Item Functioning
EFA	Exploratory Factor Analysis
FSIQ	Full Scale Intelligence Quotient
g	General Intelligence
Gc	Crystallised Ability
GCSE	General Certificate of Secondary Education
Gf	Fluid Ability
Gs	Perceptual Speed
Gv	Spatial Visualisation
ICC	Item Characteristic Curve
IQ	Intelligence Quotient
IRT	Item Response Theory
MG-CFA	Multiple Group Confirmatory Factor Analysis
MIMIC	Multiple Indicator Multiple Causes
ML	Maximum Likelihood Estimator
NNAT	Naglieri Nonverbal Ability Test
PCA	Principal Components Analysis
RMSEA	Root Mean Square Error of Approximation
SAT	Scholastic Aptitude Test
SEM	Structural Equation Modelling
SES	Socio-Economic Status

Acronym	*Definition*
SPM+	Standard Progressive Matrices Plus
SRMR	Standardised Root Mean Square Residual
TLI	Tucker-Lewis Index
WAIS	Wechsler Adult Intelligence Scale
WISC	Wechsler Intelligence Scale for Children
WLSMV	Weighted Least Means Squares Estimator
WRMR	Weighted Root Mean Square Residual

— 1 —

An Introduction to Intelligence: Definition, Measurement, & Controversy

"The answer to the question of which is the smarter sex depends on how 'smart' is defined"
(p. 230, Halpern & LaMay, 2000).

1.1. Introduction

Sex differences in intelligence continues to captivate psychologists and the general public alike, as evidenced by the ever-increasing collection of empirical research, books, informed commentary, and popular culture. No other concept in psychology has generated more debate (Johnson, 2004), and may arguably be the longest-running and most impassioned controversy in psychology's history (Halpern, In Press). This dissertation contributes to the debate by providing a psychometrically robust evaluation of sex differences in general intelligence as measured by the U.K. standardisation edition of the Raven's Standard Progressive Matrices Plus (SPM+; Raven, Court, & Raven, 2008). The representativeness of the U.K. population, the size and recency of this sample and robustness of analyses are novel contributions to the literature.

The very nature of intelligence lends itself to debate. The subject of intelligence is the most studied and likely the most understood subject in psychology, yet there remain many "unknowns" (Gottfredson & Saklofske, 2009). The field is rife with disagreement over a number of central issues at its very foundation: conflicting definitions and theoretical perspectives, disagreement over measurement practices and methodology, and questionable research practices and conclusions. Before reviewing the literature on sex differences in general intelligence, the field of intelligence will first be introduced. In order to effectively discuss group differences in intelligence, it is important to first establish some general understandings of the field of intelligence: its origins, how it is defined, and the conceptual issues that lie therein. A brief, overview of the relevant theory and literature will be provided, acknowledging, but not aligning to, different theoretical perspectives. In so doing, particular attention will be paid to the contentious issues mentioned above, by addressing the following questions:

1. What is intelligence?
2. How is intelligence measured?
3. What is the structure of intelligence? Is there one type of intelligence or many?
4. Are IQ scores rising over time?

By addressing these questions, the discussion of sex differences will be placed into context of the field as a whole. This will allow for more insightful understanding in subsequent chapters of the underlying issues as to why the subject of intelligence and sex differences is still being so fiercely debated and still worthy of investigation.

1.2. INTELLIGENCE

Intelligence, as a concept, has been with us for millennia. The theoretical concept of intelligence and the comparison of individuals with respect to their intellect was recognized and noted as early as 6th Century B.C. by one of the most ancient of Greek writers, Homer: "So true is it that the Gods do not grace all men alike in speech, person, and understanding" (Homer, 2007, p. 63).

Since that time, academics and laymen alike have regarded intelligence as a topic of considerable interest (Cianciolo & Sternberg, 2004), and remains one of the most highly regarded personal attributes, second only to good health (Gottfredson, 1998). It also remains a topic of intense academic focus, evidenced by the vast collection of literature, empirical research and commentary on the subject that continues to amass in present day. A recent search on Google Scholar generated 1,570,000 results while the more empirically-focused PsycInfo database generated 155,555 results. These vast numbers provide a quantitative indication of just how much people want to know and share what they know about intelligence. There is likely to more written about intelligence than any other subject in psychology (Johnson, Carothers, & Deary, 2008).

Like other topics of great social interest, it is not without its controversies. At the centre of the debate are methodological issues, specifically how best to define and measure intelligence (Halpern & LaMay, 2000). The concept of intelligence is different according to ones' own theoretical perspective, inherent biases, and conceptions with seemingly as many definitions of intelligence as there are investigators of it (Sternberg, 1985).

1.2.1. What is Intelligence?

The seemingly simple question – 'What is intelligence?' – has generated much debate and discussion in the search for a conclusive answer (Mackintosh, 2001; Neisser, Boodoo, Bouchard Jr., et al., 1996) and the search continues. Throughout the literature, intelligence is defined in a number of ways, each slight variations of the next: "Innate general cognitive ability" (p. 187, Burt, 1955); "the ability to understand complex ideas, to adapt effectively to the environment, to learn from experience, to engage in various forms of reasoning, to overcome obstacles by taking thought" (p. 77, Neisser, Boodoo, Bouchard Jr., et al., 1996); "The aggregate or global capacity of the individual to act purposefully, to think rationally, and to deal effectively with his environment" (p. 3, Wechsler, 1944); "The mental capacity of emitting contextually appropriate behaviour at those regions in the experiential continuum that involve response to novelty or automatisation of information processing as a function of meta-components, performance components, and knowledge-acquisition components" (p. 319, Sternberg, 1985); "Superior understanding; quickness of mental apprehension, sagacity" (Oxford University Press, 1999).

What these different definitions illustrate is that, as a concept, there is certainly a general understanding of what is meant by the term "intelligence";

it is something that everyone knows and understands at some level, yet is elusive to finding a concrete definition that everyone would agree upon wholeheartedly. The classic tale of three blind men and an elephant is a lovely illustration of the varying conceptions of intelligence.

Never having encountered an elephant previously, three men had different impressions after each touching a different part of the animal's body. The first man touched the trunk and thought the elephant was snake-like; the second man touched the leg and believed it to be like a tree; the third man felt the elephant's side and believed it to be like a wall. Of these three differing interpretations of an elephant, which was correct – all or none? Like the blind men, those who currently study intelligence cannot see what is being studied (Sternberg, 1990). In "feeling" and "exploring" intelligence in different ways, different theorists arrive at different definitions. While some might argue that differing conceptions of the same issue are cause for concern, Neisser et al. (1996a) argue that differing conceptions of intelligence need not be seen as problematic, but rather an opportunity for greater understanding.

In an attempt to further understand the nature of the differing definitions, a number of psychologists asked 2 groups of people, one group of psychologists who studied human intelligence, and another group of non-psychologists, to describe characteristics of an intelligent person. The responses from the 2 groups were not identical but were qualitatively similar, which can be classified into three categories of intellectual functioning: problem-solving ability, verbal intelligence, and practical intelligence or social competence (Sternberg, 1982).

What is evident from the first two categories is that, for many, intelligence is conceived of as those abilities and skills that one needs to be successful in formal education (such as reading, writing, and mathematics). However, as evidenced by the third category, others would argue that there are other types of intelligence that are more difficult to assess: creativity, social skills, physical and kinaesthetic. Sternberg's Triarchic Theory (1985) and Gardner's Theory of Multiple Intelligences (1993) are comprised of multiple abilities to conceptualise intelligence of the individual as a whole. While both theories have made important contributions to the field of intelligence research, the focus of this dissertation will remain on the perspective of intelligence that is cognitively-based and psychometrically testable.

1.2.2. How is Intelligence Measured?

From the time of the ancient Greeks, there has been the implicit understanding of the mental abilities of an individual. However, it was not until the late 19th Century that attempts were made to explicitly quantify intelligence in order to better understand it. What resulted was the development of the intelligence test. Intelligence tests hold an important place in modern Western society and are widely used in many settings. They are used for diagnosis, evaluation, and selection (Neisser, Boodoo, Bouchard, et al., 1996), and their use has changed very little from the time when they were first conceptualised.

Sir Francis Galton

The original concept of assessing an individual's intelligence is most often attributed to Sir Francis Galton. While most of his work focused upon the evaluation of sensory functioning (such as "keenness of sight and hearing", "breathing power", and "force of blow"; p.245, Galton, 1908), his underlying belief was that inherited differences in sensory discrimination ability were positively correlated with intellectual ability (Brody, 1992). Although the methods proposed by Galton for the assessment of intellectual functioning were later discredited, it was his influence that inspired future generations of psychologists, such as Alfred Binet and Théodore Simon.

Alfred Binet

The French Ministry of Public Instruction charged Alfred Binet with the task of devising a diagnostic instrument to quickly and reliably assess a child's ability to be effectively educated in a normal school environment (Binet, 1905). The resulting test contained items that required the child to perform a variety of mental tasks, such as naming parts of the body, remembering list of digits, copying designs from memory, and comparing weights and measures. In assessing intellectual competence, test items were ordered according to the age at which the majority of the children in the sample could solve them. This allowed him to determine whether a child's test performance was average, advanced or delayed relative to his peers and in so doing, established the model upon which the majority of psychometric measures are now based - test scores of the individual are compared to the normative reference group to assess their relative fit (Mackintosh, 2001).

Modern Psychometric Testing

While the test devised by Binet and Simon (1905) set the standard at the time, it was the considerable number of revision and adaptations that followed that proved to make significant contributions to the field of psychometrics. These are notably the Stanford-Binet (Terman, 1916), and two of the most important and widely used measures currently in use: the Wechsler Scales of Intelligence (Wechsler, 1997; 2003) and the Raven's Progressive Matrices (Raven, Raven, & Court, 1998a). In developing the Stanford-Binet Intelligence Scale, Terman obtained much more accurate information about normative age-related results allowing him to establish the definition of IQ: (mental age / chronological age) x 100[1].

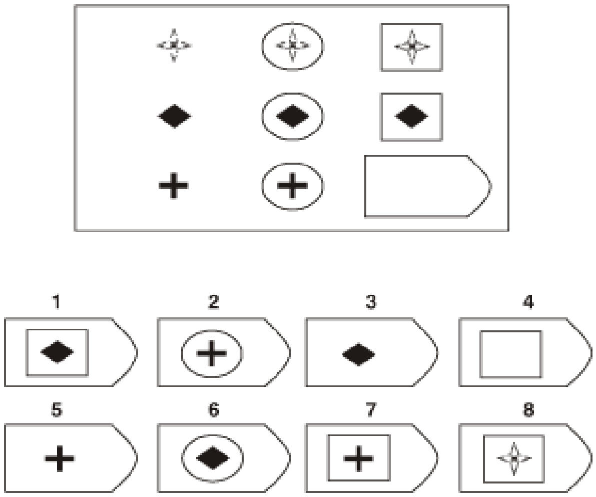
Just as with the Stanford-Binet, the Wechsler Intelligence Scales for Children (WISC) and for Adults (WAIS) were designed to measure different aspects of cognitive functioning. They are individually administered clinical instruments consisting of 15 subtests such as Vocabulary (where the participant must provide definitions of presented words), Picture Completion (where the participant is asked to identify the missing portion of an image within a time limit), and Block Design (where the participant is asked to re-create a design using coloured blocks within a time limit). The subtests are thought to measure different aspects of intellectual functioning: verbal comprehension, perceptual reasoning, working memory, and processing speed. Together they provide a Full Scale Intelligence Quotient (FSIQ) which is the modern conception of IQ defined by Terman (above).

In contrast to the multiple subtest structure of the WISC and the WAIS, the Raven's Matrices was designed with one measurement objective:

[1] Although it is noted that this definition is no longer commonly used.

to assess general cognitive ability (Raven, 2009). The Matrices were constructed as a measure of the educative component of *g* (Raven, Court, & Raven, 2008), which is the ability to forge new insights, to discern meaning in confusion, to perceive, and to identify relationships (Spearman, 1927). Due to its very clear measurement objectives, the Matrices are often used in investigations of sex differences in general intelligence, or *g* (Raven, Court, & Raven, 2008; Lynn, Allik, & Irwing, 2004). The Matrices are comprised of one type of item that assesses analogical reasoning and completion tasks (Figure 1). The examinee is presented with a 3 x 3 matrix of diagrams, where the 3rd diagram in the 3rd row missing.

Figure 1. Example of an item similar to those in the Raven's Progressive Matrices (Costa, Azambuja, Portuguez, & Costa, 2004).



The examinee is asked to choose, from a number of alternatives, the item that will complete the 3rd row. A discussion of sex differences as measured by the Raven's Matrices will be provided in Chapter 6, while further technical information of the measure will be provided in Chapter 3.

Despite being designed as a unidimensional assessment of general cognitive ability (Raven, 2009), there continues to be considerable debate in the literature about what the Raven's Progressive Matrices is actually measuring: one factor of general intelligence (Jensen, 1998; Spearman, 1927) or multiple factors of different abilities (Gustafsson, 1984; Lynn, Allik, & Irwing, 2004; van der Ven & Ellis, 2000). Further discussion of the literature pertaining to the factor structure of the Raven's Matrices will be provided in Chapter 2.

Another controversy central to the field of intelligence is highlighted in considering different scales of intelligence based on contrasting theoretical structures (i.e., the Wechsler scales versus the Raven's Matrices), and different theoretical perspectives on the same scale of intelligence (i.e., the numerous ways to view the Raven's Progressive Matrices). Just as there are a number of ways to define intelligence, and different ways to measure intelligence, there are also many ways to conceptually organise what the intelligence tests are measuring.

1.3. The Structure of Intelligence: One factor of intelligence or many?

In assessing intelligence through the use of measures such as the Wechsler scales and the Raven's Matrices, psychometricians have identified a wide range of cognitive abilities[2] that are conceptually distinct from one another, and yet statistically related, most often determined through the application of factor analysis.

When considering a collection of related indicators (such as a series of items on an intelligence test), factor analysis techniques serve to understand the variation and covariation, or patterns of relationship, among a collection of test items in the simplest, most parsimonious manner (Brown, 2006). The resulting factor structure provides suggestions about underlying causes of the covariation.

Explanations of the factor structure of intelligence can vary between theorists depending on their underlying theoretical conceptions of intelligence. Herein lays the "fuel" that fires the debate about the structure of intelligence that has existed since the time when intelligence began to be measured.

No other concept of psychology has generated more debate and may

[2] Cognitive abilities are understood to be theoretical constructs that represent the underlying components of intelligence (Halpern, 2000).

arguably be psychology's "longest-running and most acrimonious controversy" (p. 24, Halpern, In Press). Some theorists argue that intelligence comprises a number of separate elements of intelligence, such as verbal, spatial or analytic abilities (Gardner, 1993; Thurstone, 1931; Thurstone, & Thurstone, 1941). A conception of intelligence comprised of multiple abilities is often referred to as "modularity" (Halpern, In Press). Others focus on the variance that all such abilities have in common, what is commonly known as general intelligence or *g* (Spearman, 1927). This unitary view of intelligence is often referred to as "molarity". Others still conceive of intelligence as a combination of these two approaches, in something of a hierarchical arrangement of multiple factors, with an overarching general factor, *g*, at the top (Carroll, 1993). These will now be discussed.

1.3.1. One Factor of General Intelligence – g

In the early study of intelligence, it was noted that people who performed well on one kind of mental test were likely to perform well on others (and similarly for those who performed poorly; Demetriou, Mouyi, & Spanoudis, 2008; Gottfredson, 1998). This commonality suggested some universal element that has come to be known in the literature as 'general intelligence' or *g*.

This universal concept of general intelligence was first statistically described by Spearman (1927) after determining, through the application of principle components analysis, that much of the variability in people's intelligence test scores could be attributed to one single, common factor. His principle of the *indifference of the indicator* countered the argument made by Thurstone that an understanding of intelligence was intrinsically linked to the test content. Rather, Spearman believed that the only concern was item *g*-loadings: because all IQ tests correlate highly with one another, a general factor extracted from one test would be the same as one extracted from another measure.

Although Spearman acknowledged factor analysis, and *g*, as the description of a pattern of interrelationships among any number of different intelligence tests, he also acknowledged that this description is not the same as identifying human abilities (Mackintosh, 2001). Rather, he saw the pattern of interrelationships as indicative of an underlying psychological process, a type of mental energy (Howe, 2000).

This conceptualisation of *g* continues to be another contentious point of disagreement in the intelligence literature. To some, *g* is described in such a way that suggests that it is a tangible entity, an “active ingredient” (p. 34, Rushton, 1995) that someone either has or doesn’t have in a finite amount. This conceptualisation has been vehemently disputed with “no convincing reason...for insisting that *g* is something real” (Howe, 2000, p. 30). *g* is simply a means for statistically describing commonalities among a series of test items designed to assess cognitive functioning. By conceptualising *g as* something one has or doesn’t have runs the risk of sentencing individuals into categories of ability, rather than measuring individuals for their cognitive potential.

A closer look at Spearman’s theory reveals the finer points of his notion of the general factor of intelligence: that there are, in fact, two factors (Colom & García-López, 2002; Mackintosh, 2001). The general factor, as previously described, that accounts for the correlations among different tests as well as a secondary specific sub-factor unique to each test. Together they form “the two-factor theory of intelligence” (Spearman, 1927). He described the specific factors (or *s*) as those processes needed to solve a specific type of problem, or “nuts and bolts” of intelligence, while the energy to drive to overall “intelligence engine” came from the general factor (Mackintosh, 2001).

This discussion of variations of interpretations highlights the importance with which results of intelligence research ought to be considered – with caution and with awareness of how they fit within the larger context of social understanding. Factor analysis can simply describe relationships among different sets of items and different tests, which is distinctly different to determining the structure of human abilities (Mackintosh, 2001). A relationship between constructs does not necessarily imply cause and effect.

1.3.2. Multiple Factors of Intelligence

Many theorists conceptualise intelligence as a complement of several different intellectual abilities (Halpern, n d)). Drawing a parallel to the variety of psychometric tests that are available to assess intelligence, the multiple factor perspective would be akin to the Wechsler scales, measuring different abilities in separate subtests. But even among those theorists who believe intelligence is comprised of different abilities, they do not agree on how many different abilities there are.

Moving on from Spearman's findings of *g* and *s*, Cattell (1970) believed there to be more than one second-order factor. He proposed that *g* could be divided into two separate sub-factors: fluid and crystallised ability. He and Horn (1966) went on to subsequently propose nine ability factors including crystallised ability (Gc), the accumulation of knowledge and skills, and fluid ability (Gf), the capability of abstract reasoning and flexibility of thought. These were then further fractioned into spatial visualisation (Gv) and perceptual speed (Gs) among others (Mackintosh, 2001). Carroll (1993) suggests that it will be infinitely possible to further decompose these broad factors into more specific, subsidiary factors.

Another model that continues to receive wide support is that proposed by Thurstone and Thurstone (1941). They believed that the content of test items intrinsically influenced the outcome of factor analysis, and consequently, upon the interpretation of a theory of intellectual abilities. Their extensive factor analysis of 60 different intelligence tests revealed three categories of abilities: verbal, number (quantitative), and perception (visual-spatial). These three factors of intelligence are still widely referred to in the current literature.

Another particularly influential model of multiple factors of intelligence is Gardner's Theory of Multiple Intelligences (1993). In his book *Frames of Mind*, Gardner contested that traditional IQ tests only measure a sub-set of abilities, and to effectively assess an individual's overall level of ability, it is important to account for other non-traditional aspects of intelligent functioning. In total, Gardner has proposed seven different intelligences: linguistic, spatial, logical-mathematical, musical, bodily-kinaesthetic, interpersonal, and intrapersonal. Each of these 'frames of mind' was arrived at through different sources of evidence, not only psychometric evaluations.

Not all intelligence experts agree with Gardner's definitions of alternative abilities, and some would argue that according to Gardner's criteria for identifying intelligence, a far greater range of abilities would be possible with virtually any specific skill or cognitive operation qualifying to be defined as an 'intelligence' (Mackintosh, 2001; Brody, 1992; Sternberg, 1990).

While they differ in conceptualisation and configuration of their factors, the notion of multiple abilities in intelligence is still widely accepted and further discussion will be provided in relation to sex differences in Chapter 2. While theories of multiple factors of intelligence (such as those proposed by Thurstone & Thurstone, [1941] and Gardner, [1993]) continue to make important contributions to the field of intelligence research, the focus of the current dissertation will remain aligned with the conceptual framework of the Raven's Progressive Matrices, and hence, upon a one-factor model of general intelligence.

1.3.3. Multiple Factors of Intelligence and g – A Hierarchy

An alternative way to view intelligence combines the single factor and multiple factor perspectives just described, in something of a compromise. Rather than a collection of multiple unrelated abilities or one single *g*, intelligence can be conceptualised as a hierarchy of different abilities with *g* overarching them all. While acknowledging the many meaningful separate factors exist, an extensive literature review by Brody (1992), concluded that "the structure of abilities tests supports a hierarchical model of ability with g at its apex" (p.40).

It could be argued that by describing intelligence in this hierarchical configuration allows for 'the best of both worlds': the specificity and attention to detail of the multiple factor perspective with the parsimony and simplicity of the general factor model. The fact remains that performance on one kind of intelligence test correlates highly with performance on all other kinds of tests, and due to this "positive manifold" (Spearman, 1927), a large general factor will always be obtained through factor analytic investigations (Mackintosh, 1995; 2001).

Despite the differing theoretical perspectives on the structure of intelligence, the strength of g continues to be supported. It is upon the strength of the general factor that this dissertation is established.

1.3.4. Are IQ scores increasing? The Flynn Effect

Another issue that has arisen in the intelligence literature in recent years pertains to the worldwide generational increase in IQ scores over time – known as the Flynn Effect. James Flynn (1987) identified that, in the last 50 years, IQ raw scores have increased by more than 15 points (or approximately 3 IQ points per decade). The greatest gains appear to be on tests that are designed to be free of cultural influence, such as the RPM, and the rate of gain may be increasing.

Reasons for this increase are unclear. One explanation for these gains has often been the increasing sophistication in measurement practices. Intelligence tests are often ‘re-standardised’ in order to account for increasing population intelligence (Neisser, Boodoo, Bouchard Jr., et al., 1996). As part of the process, the mean score of the new standardization sample is recalibrated to 100. In doing so, increases in population IQ are often masked. This means that if 20 years had elapsed between standardisation editions, a group of people tested on both versions would, for example, score 106 on the older version but 100 on the newer version. Some would argue that the steady increase in IQ scores described by Flynn is too large to result from increased test sophistication and improved testing practices (Neisser, Boodoo, Bouchard Jr., et al., 1996), and they must therefore be attributable to the abilities of the test takers themselves.

Improvements in nutrition have been offered as another explanation (e.g., Roberto Colom, Lluis-Font, & Andres-Pueyo, 2005). (Lynn, 1990) noted large nutritionally-based increases in height during the same period as the IQ gains, and Neisser et al. (1996a) question whether nutrition might also have increased brain size, and therefore are responsible for the rise in IQ. They do note that a clear causal relationship between nutrition and intelligence has yet to be conclusively established.

Flynn (2009) refutes the relationship between nutrition and IQ gains, claiming that the two trends are largely independent on strength of multiple factor analyses of numerous standardisations of the Raven’s Standard Progressive and Coloured Matrices. Flynn himself says “Enhanced nutrition has made us taller people and poorer nutrition has made us more obese. But our diet today probably does not make us very different people from our grandparents as far as cognitive competence is concerned" (p. 26, 2009). Further he maintains that increases in IQ do not necessarily indicate an increase in intelligence (personal communication, October 11, 2010).

Rather, Flynn (2009) attributes the gains to the Industrial Revolution and increasingly sophisticated educational practices. There is no question that there are notable cultural differences between successive generations with respect to the accumulation of information and knowledge. Daily life and occupational experience seem more complex today than ever before. For example, The New York Times contains more information in one week day edition than the average person was likely to come across in a lifetime in 17th century England (Wurman, 2000). Populations are becoming increasingly urbanised, children stay in school longer than ever before, and almost everyone seems to be encountering new forms of cognitively complex experiences through television and the internet. Through such experiences, cognition has developed accordingly, and is qualitatively different to that of previous generations resulting in a group of people who are better at solving the problems that are relevant to them (Flynn, 2009).

Another possible explanation of the IQ gains of the Flynn Effect is measurement bias that arises from something other than the latent construct of intelligence. One such explanation might be Differential Item Functioning (DIF). DIF can occur when people of the same level of intelligence from different groups (such as different generations or sexes) have a different likelihood of giving a correct answer on an intelligence test due to something inherent in the test item itself. For example, within the context of increased exposure to technology, different generations will likely have different exposure to completing spatial tasks that are common in video games (Feng, Spence, & Pratt, 2007; Subrahmanyam & Greenfield, 1994). Test items, like those on the RPM that assess spatial ability, will likely be easier for current generations not because they are more intelligent, but because the current generation will have considerably more experience with spatial tasks like these. Consequently, these items are biased against members of previous generations. Through the advent of modern statistical techniques such as Structural Equation Modelling (SEM) and Multiple Groups

Confirmatory Factor Analysis (MG-CFA), it is increasingly possible to investigate such sources of measurement bias. It is clear that the Flynn effect has yet to be understood fully from the perspectives of psychometrics (Wicherts, 2008).

With better understanding of the nuances of group differences in intelligence, whether they are groups of males and females or groups of individuals at different points in history, through the use of advanced statistical modelling techniques we will begin to gain a better understanding of human intelligence. The current dissertation will be employing SEM and MG-CFA techniques with the U.K. standardisation data of the SPM. The use of these sensitive statistical methods is unprecedented in the literature, and is considered a further strength of this dissertation.

1.4. Summary

While it is true that a great deal is known about intelligence, there is arguably still much to understand about this complex topic. What is most apparent from a review of the literature pertaining to intelligence is that there is disagreement on a number of accounts pertaining to the definition of the construct, the method of measurement and its structure. The extent of the debate and lack of consensus needn't be considered an obstacle, as discussed by Neisser et al. (1996), but rather as an opportunity to learn more about intelligence.

When prominent theorists were asked to provide a definition of intelligence, each gave a different definition (Sternberg & Detterman, 1986). For some psychologists, intelligence is considered a collection of many different abilities. "Individuals differ from one another in their ability to understand complex ideas, to adapt effectively to the environment, to learn from experience, to engage in various forms of reasoning, to overcome obstacles by taking thought" (p. 77, Neisser et al., 1996). From a different perspective, separate intellectual abilities are found to be positively correlated with one another, leading many psychologists to view intelligence as a single, general construct, *g.*

Equally contentious is the debate of measurement. Since the time of Binet and Galton, there has been a debate about the best way to measure intelligence. Some would argue that "there is no single true measure of intelligence because there is no single thing called intelligence" (p. 582, Mackintosh, 1995), as evidenced by Gardner's Theory of Multiple Intelligences. For those who conceptualise intelligence in terms of multiple cognitive abilities, a multiple subtest approach is most appropriate, as is seen in the WISC or the WAIS. For those who conceive of intelligence in terms of one general intelligence, *g*, tests such as the Ravens are considered to be optimal.

While there is much value in each of these perspectives, in order to present a cohesive argument throughout this dissertation, it is necessary to align with certain theoretical perspectives and not others. The main objective of this dissertation is to assess whether there exists sex differences in general intelligence using modern statistical modelling methods. As such, the focus will remain upon the unidimensional construct of general intelligence as measured by the Raven's Progressive Matrices.

In light of the review of the contentions in the field of intelligence, it is now appropriate to proceed with a comprehensive review of the relevant literature that pertains to sex differences in intelligence in Chapter 2.

— 2 —

Literature Review:

Sex Differences in Intelligence

2.1. Introduction

Since before the time when tests of intelligence were first developed by Binet (1905) and Galton (1908), there has been great interest in differences between people: what it means for one person to be more intelligent than another, and how these differences may have arisen. Herein lays the focus of this chapter.

Of the many controversial issues that are found in the literature on intelligence, “few have generated more heat and less light” (p. 559, Mackintosh, 1996) than the discussion of group differences in average IQ, particularly with respect to differences between males and females. Much has been written on the subject, and still there has yet to be a consensus reached. Despite this lack of consensus, the subject continues to gain interest with the number of publications on this topic soaring. A recent search for sex differences and intelligence on PsycINFO returned over 86,000 entries, while a search on Google Scholar returned 652,000 suggestions for empirical articles on the subject.

Throughout this extensive literature, there are distinct trends emerging from the findings. When intelligence is conceived as a collection of multiple cognitive abilities, reports indicate that there are some cognitive domains where, on average, females excel and others where males excel. While a considerable amount has been learned about intelligence in recent years, surprisingly little has changed in terms of the overall conclusions drawn from the beginning of last century.

> “The chief ascertainable differences appear to be the following: boys are better at arithmetic, mathematics, physical sciences, classical languages, geography, and drawing; girls are better at reading, spelling, handwriting, English composition, English literature, and possibly history, modern languages, and biological sciences” (p. 658, Nature, 1923).

On tasks of visuospatial ability, such as mental rotation[3], large differences have been consistently found favouring males (Linn & Petersen, 1985; Moè, 2009; Voyer, Voyer, & Bryden, 1995). On some tasks of verbal abilities, such as verbal fluency (e.g., a task of naming words that begin with a given letter in a given period of time) and synonym generation, females are found to outperform males (Hines, 1990; Hyde & Linn, 1988). Further, females are found to excel on a number of verbal measures such as reading comprehension and spelling (Hedges & Nowell, 1995), and tend to score higher on tests of literature and composition (Stanley, 1993).

When intelligence is considered as a general construct, commonly referred to as *g*, the existence of differences between the sexes continues. Many would conclude that males are the more intelligent sex overall (Abad, Colom, Rebollo, & Escorial, 2004; Jackson & Rushton, 2006; Lynn, 1998; 2002; Lynn & Irwing, 2004; Vigneau & Bors, 2008)

Some suggest that the different rates of physical and cognitive maturation of males and females influence the emergence of sex differences, resulting in a female advantage before 15 years of age and a male advantage afterwards (Lynn, 1994). Others would conclude that there are no meaningful sex differences in mean scores of general intelligence. Any differences that have been found tend to be considered “small and virtually non-existent (Brody, 1992; Colom & García-López, 2002; Halpern, 2000; Jensen, 1998; Mackintosh, 1996; Rushton & Cvorovic, 2009). For

[3] Mental rotation – the ability to rotate stimuli rapidly and accurately within the mind (p. 12, Hines, 2005).

others still, sex differences have been demonstrated in variability of scores if not in average IQ (Deary, Thorpe, Wilson, Starr, & Whalley, 2003; Hedges & Nowell, 1995; Irwing & Lynn, 2005).

Reasons for the noted sex differences in multiple abilities of intelligence and *g* alike have been explained by a number of different factors: biological (such as anatomical differences, genes and hormones) environmental (such as social influence and improved nutrition), or a combination of factors as suggested in Halpern's biopsychosocial theory (In Press). Some would argue that the differences reported in the literature are not, in fact, differences at all but rather psychometric artefacts arising from statistical irregularities such as item bias and differential item functioning (van der Sluis et al., 2008; Wicherts, Borsboom, & Dolan, 2010).

These will each be discussed in turn through a comprehensive review of the sex differences literature with special consideration for the unidimensional perspective of intelligence and the Raven's Standard Progressive Matrices (SPM; Raven, Court, & Raven, 2008).In so doing, the literature underlying the aims of this dissertation will be reviewed. The first aim of this dissertation is to determine whether the SPM+ is measuring the construct of general intelligence in the same way for both males and females. That is to say, is the SPM+ free from item bias, ensuring a fair assessment for both sexes? The second aim of the dissertation is to determine whether there is a significant sex difference in the mean and variability of scores in the overall sample of the SPM+. Due to the large age range of participants of the standardisation sample, it is also important to consider whether an overall evaluation of sex differences are masking the emergence of developmentally-related sex differences in general intelligence. This forms the third aim of this dissertation. The final aim of this dissertation is to determine whether extraneous elements inherent in the items, or method effects, are affecting performance on the SPM+.

2.2. WHY STUDY SEX DIFFERENCES?

In all areas of research in intelligence, there exists controversy about the way that research is conducted, how the results are interpreted, and the significance of the results from both theoretical and practical perspectives. For research in sex differences in intelligence, there is an additional controversy of whether the research ought to be conducted at all (Halpern, n d)). What is particularly interesting is that much of the controversy stems not from the discovery of differences themselves but from the explanations used to understand the findings. These, according to Cianciolo and Sternberg (2004) are not always based upon empirical data, but upon predetermined agenda. Some fear that the results of studying sex differences will be interpreted and used in such a way to support a misogynist objective and oppress women or to reaffirm 'traditional' roles for males and females in society.

While the potential reasons for not studying sex differences raise valid concerns, there are many more reasons why the study of the differences between males and females ought to continue. Investigations of sex differences in cognitive ability provide valuable insight into a number of areas that are deemed socially important, not only those directly associated with scores on intelligence tests and academic achievement.

Societal inequalities continue to prove detrimental for females in Britain and around the world. It is well documented that females are less likely to enter into certain careers (such as science, technology, engineering or maths; Eccles, 1994) or into senior and executive positions within business and academia (Niederle & Vesterlund, 2007). Seemingly associated with this, there continues to be a significant wage gap between males and females, particularly in legal, healthcare and technical occupations. For 2009, the median weekly pay for full-time male employees was £531 per week, while for females it was £426. Although the wage gap narrowed by 0.4% compared to 2008, there still remains a 12.2% gap in the amount of money males and females are being paid for equivalent positions of employment (National Statistics, 2009). Presuming that individuals in similar occupations have obtained similar levels of education and possess similar levels of intellectual ability, such figures suggest evidence of discrimination and bias against females.

When discussing classroom-based differences between males and females, it is the assessment of student achievement that is the focus. Tests of achievement aim to measure the amount of knowledge an individual possesses in a particular academically-based subject area (like reading or maths). Although inextricably related, it is qualitatively different to the assessment of ability. Tests of ability aim to measure an individual's proficiency in a particular domain (such as verbal fluency or spatial rotation). These underlying abilities, or latent constructs, are relatively abstract constructs, and can prove quite difficult to effectively quantify and measure.

There are many reports in the intelligence literature that document one group excelling over the other group; in the achievement literature, girls tend to outperform boys, while in the ability literature, males are often cited as outperforming females. Achievement is directly related to the education one receives in school, and is inherently entwined with ability. In practical terms, it is difficult to devise a test of ability that does not also measure an element of achievement. It is, therefore, not surprising that intelligence scores correlate very highly with school achievement and with the number of years of an education an individual will complete (Deary, Strand, Smith, & Fernandes, 2007; Mackintosh, 2001). Because it is virtually impossible to isolate ability and achievement as independent constructs, it is important to be cautious in interpretations of reports of sex differences in accordance with the measurement approach taken.

Performance differences between boys and girls in the classroom have been noted worldwide. In the UK, recent reports from the 2010 National Curriculum results for Key Stages 2[4] show that girls perform better than boys in reading and writing. It is also noted that the performance of girls has improved in mathematics and is now equivalent to boys for the first time since 2004 (Department for Education, 2010).

A similar picture emerges from the recent GCSE results published in August 2010. 72.6% of females and 65.4% of males received passing grades of A* to C. This has prompted discussion over whether different versions of the tests are needed in order to capitalise on the perceived strengths of male and female students. The *Assessment and Qualifications Alliance* claims that assessments such as this would allow for greater flexibility in order to match the specific needs and differences of boys and

[4] Reports are for children who are 11 years of age. Reports showing female advantage are also available for children of Key Stage 1 (age 7 years) and Key Stage 3 (aged 14 years) but they incorporate a teacher assessment. Teacher assessments may be influenced by extraneous variables, and therefore not deemed appropriate for the current discussion.

girls, and may be ready as early as September 2011(Collins, 2010).

Results such as these are often sensationalised in the media leading to claims that either one group or the other is being disadvantaged by the current education system. For many years, research claimed that schools and teachers were biased against girls, and that this has contributed to their underachievement relative to males in such domains as maths and sciences (American Association Of University Women, 1993). However, there has been a shift of perspective in recent years. Research findings are now pointing towards what the media is now referring to as a "boy crisis" (Gurian & Stevens, 2005) and the "*New Gender Gap"* (Conlin, 2003). "There is a rather alarming crisis in our society that's been developing for years...Too many of our boys and young men are falling behind in school and life...Something has gone wrong in the way we educate our boys" (p. 28, Gurian & Stevens, 2005). But whether the differing performance of boys and girls ought to be viewed as a "crisis" is open for interpretation.

Such differences in achievement and ability need not automatically mean that one of the groups is inferior. Neisser et al., (1996) suggest that investigations into sex differences and their results need not be cause for concern but are opportunities for greater understanding of intelligence as a whole. Despite being an opportunity for greater understanding, very little in the literature discusses sex similarities in intelligence or the non-significant findings. Some would suggest that a "file-drawer" problem is occurring (Blinkhorn, 2005). An argument could be made that in discussing differences it would also be prudent and constructive to discuss similarities as well. Hyde (2005; 2007) advocates for the gender similarities hypothesis, suggesting that males and females are much more similar than they are different. She indicates that empirical findings of similarities ought to take a more prominent place in the literature. The gender similarities hypothesis will be discussed further in section 2.4.1.

Although not directly related, the benefits of studying group differences in intelligence can be evidenced by one of the most intriguing phenomena in psychology, the Flynn effect (as discussed in Section 1.3.4.). It was only through the study of differences in IQ scores over time that it was concluded by Flynn that each successive generation is more intelligent than the last. Having a greater understanding of the nuances of intelligence and this complex phenomenon for males and females separately may well contribute towards understanding intelligence as a whole, as well as to facilitate research into helping *all* individuals achieve the maximum potential.

Before proceeding with a review of the literature, it is important to clearly establish what is meant by sex differences in this dissertation. Some authors attempt to differentiate between sex differences and gender differences, where sex differences are understood to be biologically based and gender differences are socially determined. However, what is genetically or socially determined is extremely difficult to distinguish. Regardless of the biological or social determinants of the construct under discussion, the term “sex differences” is used throughout this dissertation in order to facilitate a clearer conceptual construct for the reader.

It is also important to state that within the literature, the concept of sex differences is referring to an average difference between the sexes not an absolute one. No individual is average, and as such, group averages tell us nothing of the individual. For example, if discussing sex differences in weight, it does not imply that all men are heavier than women, but as a group, on average, men are heavier than women.

Within the context of this dissertation, sex differences will be understood as the *average* differences for groups of males and females. These will be represented as effect sizes, *d*, a measure of the magnitude of difference between the two groups. Effect sizes are determined by calculating the difference between the mean scores and dividing by the standard deviation Equation 1:

$$d = \frac{\overline{X_M} - \overline{X_F}}{\sqrt{\frac{SD_M{}^2 + SD_F{}^2}{2}}}$$

Equation 1: Cohen's *d* for the calculation of effect sizes (Cohen, 1988)

Effect sizes are particularly useful in a literature that employs a number of different measures and metrics for reporting scores. They provide a standardised way of assessing and comparing results and will facilitate understanding as the literature on sex differences is reviewed. The interpretation of effect sizes is dependent upon on the context of the research and is comparative in nature (Hedges, 2008). However, it is generally understood that in the behavioural sciences an effect size of 0.2 to 0.3 is considered small, an effect size of 0.5 is considered a medium effect, while an effect size of greater than 0.8 is considered large (Cohen, 1988).

2.3. Sex differences in Review

As mentioned previously, the literature on sex differences in intelligence is extensive. While the focus of this dissertation remains general intelligence as measured by the SPM+, it is useful to understand the nature of the differences reported in the literature on constructs such as spatial and verbal abilities. In addition to general intelligence, some researchers contend that the Raven's Matrices also measure aspects of spatial and verbal abilities. *g* can be thought of as a higher-order factor of intelligence that over-arches other cognitive abilities, and for this reason, a broad understanding of the domain specific differences will help contextualise the discussion of sex differences in all aspects of intelligence*.* A brief overview will now be provided of the sex differences in the domains of spatial and verbal abilities. This will be followed by a more comprehensive review of the literature relating to sex differences in general intelligence as measured by the Raven's matrices.

2.3.1. Sex Differences in Multiple Abilities of Intelligence

In the literature that approaches intelligence from the perspective of multiple cognitive abilities, significant differences between males and females have been noted on a number of constructs. It is the group differences on these specific abilities of intelligence that have attracted the most attention (Mackintosh, 2001). Throughout the literature, evidence has suggested some consistent sex differences in cognitive abilities, namely spatial and verbal abilities. Although there is considerable evidence to support such differences, there has yet to be unanimous support in the literature, as there always seem to be two sides to each story.

2.3.1.1. Spatial Ability

Spatial abilities are generally thought to be those skills required to represent, transform, generate or recall symbolic, non-linguistic information (Linn & Petersen, 1985). The sex differences in spatial ability are considered to be among the largest of differences acknowledged in the literature of all the cognitive abilities (Lawton & Morrin, 1999), but the magnitude of the difference varies according to the type of ability measured and the measurement method employed.

One example of a spatial ability task, mental rotation, requires the individual to imagine what an object would look like if it were rotated, for example, 90 degrees in a clockwise direction. Another example would be the Water-Level test, where the individual is asked to determine what the level of water would look like when a bottle is tilted to different orientations (see Figure 2). On tasks such as these, males have been found to consistently outperform females.

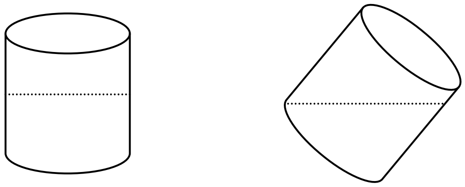


Figure 2. Water-level Test

In a seminal work, Maccoby and Jacklin (1974) established that sex differences in spatial abilities exist in favour of males (d_{avg} = 0.45). One particular critique of their work was that they were inconsistent with the criteria used to decide if the proportion of studies indicating a sex difference was sufficiently large to conclude the existence to a true sex difference (Block, 1976). Further, they did not provide precise measurements of magnitude of the sex differences for the different types of tasks that were being assessed. Further investigation by Linn & Petersen (1985) determined that spatial ability is not a wholly universal concept. They determined three factors of spatial ability: spatial perception, mental rotation, and spatial visualisation[5], again with males outperforming females on each of these factors.

The validity of such meta-analytic procedures with spatial ability data has been questioned (Caplan, MacPherson, & Tobin, 1985), due to the argument that there is little agreement about the true nature of spatial abilities. Yet, meta-analysis appears frequently in the literature as a way to summarise the findings in this prolific area. The critique by Caplan et al. was addressed in a further meta-analytic study by Voyer, Voyer, and Bryden (1995). Their analysis of 286 studies revealed a mean weighted *d* of 0.37, which demonstrates a highly significant male advantage in overall spatial abilities (from a heterogeneous pool of spatial tasks). When they partitioned the studies into homogeneous groups of tests a similar result was noted. For example, the Mental Rotations Test yielded a mean weighted *d* of 0.67, for the Embedded Figures Task a mean weighted *d* of 0.18, suggesting that different tests are measuring different aspects of spatial abilities.

In a review of 46 meta-analyses, Hyde (2005) provided evidence that differences between males and females were found to be moderate or large on varying types of spatial ability, with the greatest magnitude of difference on tasks of mental rotation (d = 0.92). With an overwhelming body of evidence of multiple meta-analyses, it is tempting to take these findings as conclusive. However, are differences such as these inevitable? Will girls, on average, always underperform relative to boys? Why do these differences occur?

In a recent intervention study with young children, Tzuriel and Egozi (2010) administered mental rotations tasks to boys and girls before and after they participated in a programme designed to promote the representation and transformation of spatial information. They found that boys performed significantly better than girls on spatial relations tasks (d = 0.45) only in the pre-intervention phase. They concluded that their intervention eliminated the sex difference, and actually reversed it: in the post-intervention assessment, girls performed significantly better than boys (d = 0.23). These findings are

[5] Some would also suggest that there are a further two factors of spatial ability: spatiotemporal ability and the generation and maintenance of a spatial image.

with respect to very young school-aged children, immediately after the intervention, and the results may not endure. Further, these results may not generalise to older children or adults. The strength of success of the intervention might suggest that by promoting representation and transformation of spatial information, a skill not traditionally taught in classrooms (Webb, Lubinski, & Benbow, 2007), could be a means for further equalising the performance differences in later years of education.

A different, but related, study tested the motivational aspect of performance in mental rotation (Moè, 2009). Study participants completed a listening task pertaining to the capability of the groups to rotate objects in their minds. Participants were assigned to one of three conditions and were informed about the experimental task according to one of the following: men were better at the experimental task than women; women were better than men at the task; and a control task where neither gender was referenced. Participants were then asked to complete either an easy or difficult mental rotation task. In general, the results reveal that the male participants were primarily influenced by the information relating to task difficulty. The performance of females was affected by positive instructions about gender, and their performance matched that of males when they were led to believe that females were better than men on the task. As with the study by Tzuriel and Egozi, the sample was very restricted: 152 high school students ranging in age from 15 to 22 years. What remains to be seen in future replications of such studies is the longevity and wider scope of results possible with a larger representative sample.

A further study associated with stereo-type threat was conducted to assess whether children's socio-economic status (SES) and beliefs regarding differences in ability were associated with performance on the Raven's SPM (Désert, Préaux, & Jund, 2009). Stereo-type threat is the phenomenon whereby the intellectual performance of a particular group is negatively influenced by the evaluative pressure of a negative stereotype. The association between socio-economic status and educational outcome has been established in the literature (Neisser et al., 1996). However, it is yet unclear which mechanisms are involved.

In order to better understand the role of negative reputations regarding intellectual ability and level of SES, Désert et al., administered the

SPM to 153 children (mean age = 7.82 years) from 8 elementary schools in France. Two different test instruction conditions were administered to the children who were also classified according to either high or low socio-economic level. The evaluative instructions conveyed to the children that they were being assessed in terms of their intellectual performance, while the non-evaluative condition presented the SPM in terms of games and tasks that children of their age could complete. After the completion of the SPM, the children were asked three questions relating to self-evaluation of scholastic ability and general beliefs about scholastic abilities of low and high SES children.

The results of the study provided evidence that children from the disadvantaged group were influenced by stereotype threat, with the performance of children of low SES significantly lower in an evaluative condition than non-evaluative condition. In concert with the questionnaire data, the results provide evidence that social stereotypes are understood and internalised in children as young as 7 years of age, in their belief that children of higher SES are intellectually superior then children of lower SES.

Due to the culturally-specific nature of this sample, such results cannot be widely generalised. Social-stereotypes associated with economic status likely vary between countries, and therefore, replication of such a study is necessary to better understand the role of stereo-type threat in performance on the SPM. Further, in relation to this dissertation, the data from this study was not considered in terms of the gender of the participants, but in light of the potential for a gender-based stereotype to influence the performance of children, it would be important for such results to be replicated to account for gender as well as SES.

2.3.1.2. Verbal Ability

Another construct of ability where consistent sex differences have been reported in the literature is verbal abilities. Socially, the sex differences in verbal abilities have become commonly understood facts about how boys and girls are different. As with other constructs of cognition, there are many different tasks that can be classified as verbal abilities: spelling, reading, writing, word fluency, and verbal comprehension. Differences between boys and girls are generally noted from the time that girls begin to acquire the ability to verbalise and communicate vocally. Girls tend to learn to speak earlier than boys do (Ozcaliskan & Goldin-Meadow, 2005), they have larger vocabularies earlier (Lutchmaya, Baron-Cohen, & Raggatt, 2002), and tend to use more complex sentence structures at an earlier age than their male counterparts (Horgan, 1975). Although the sex difference in these skills tend to equalise as children age and develop , if a sex difference is noted in later assessments of verbal ability, the advantage is often attributed to females.

The variation in size and reliability of the sex differences noted in the literature pertaining to verbal ability reflects the constructs measured. The lack of consensus in claims about the size and specificity of sex differences in verbal ability across studies is likely the result of analysing overall data with a failure to differentiate between the different verbal tasks measured, such as verbal fluency, reading comprehension and vocabulary.

The work of Maccoby & Jacklin (1974) provided a landmark contribution to the field, with their review of verbal ability studies published between 1964 and 1974. They concluded there was evidence of female advantage of verbal skills, particularly from the age of 16 onwards. Their work also served to generate further interest in a meta-analytic approach to the study of sex differences.

Hyde and Linn (1988) conducted a meta-analysis of studies that included of a range of different tasks of verbal ability. When they considered verbal ability as an overall construct, they determined that 27% of the studies showed that females performed better than males, 7% of studies showed that males performed better than females while 66% found no significant differences. Despite the very large collective sample size (N = 1,418,999), they consider the magnitude of the sex difference attributable to verbal ability “so small that it can effectively be considered zero” (p. 64). This overall conclusion of a nearly null difference made a substantial contribution to the literature.

The landmark contribution by Hyde and Linn illustrates quite poignantly how looking at average, overall differences (or lack thereof) can sometimes mask what is going on at a more detailed level. When they assessed group differences using the same data with respect to task type, a different picture of sex differences emerged. They determined that females had greater ability in word fluency (0.33*d*) and anagrams (0.22 *d*), while men have greater ability in analogies (0.16 *d*). Negligible differences were noted in reading comprehension, vocabulary, and essay writing. Such sex differences have not been reported consistently throughout the literature, varying by task type. Results from the literature will now be briefly reviewed in terms of the type of task measured.

One task that often shows the largest consistent sex difference is verbal fluency. Tasks of verbal fluency require participants to generate as many words as possible in a specified time period. In a lexical condition of the task, the words must begin with a particular letter (e.g., the letter G), whereas in the categorical condition, the words must come from a specific category (e.g., types of animals). In a study of 97 students studying undergraduate psychology and medicine, Weiss et al., (2003) determined that females perform significantly better on lexical fluency (*d* = 0.45) but not on categorical fluency (*d* = 0.24). Such research claims cannot be considered particularly generalisable in light of the sample they used: university students studying psychology and medicine are likely to be in the upper range of the distribution of intellectual function, and are unlikely to represent average males and females.

In another study of lexical and categorical fluency, Tombaugh, Kozak, and Rees (1999) assessed 1,300 English-speaking participants ranging in age from 16 to 95 years. Accounting for both the age and education level of their participants, their regression analyses indicated that gender only accounted for 1% of the variance, leading them to conclude that there were no significant differences between males and females.

Thinking of verbal fluency from a more anecdotal perspective, females are often stereotyped as being much more loquacious than males. However, this was not found to be the case in a study by Mehl et al. (2007), with both sexes uttering approximately 16,000 words per day.

Differences with respect to reading show particularly salient female advantages (Logan & Johnston, 2010). In their secondary analysis of six large data sets, Hedges & Nowell (1995) concluded that on average, females had a "slight tendency" to perform better than males on tests of reading comprehension (with small effect sizes ranging from *d* = -0.15 to 0.002). As early as entry into school, boys are found to underperform in reading relative to girls (Tach & Farkas, 2006), and continue to underperform throughout primary school (Trzesniewski, et al., 2006). These differences are evidenced internationally as reported in Mau & Lynn (2000), Lietz (2006), and Lynn & Mikk, (2009).

In a similar task, verbal reasoning, Strand et al., (2006) found that girls outperformed boys on the verbal reasoning portion of the Cognitive Abilities Test (CAT) in a large representative UK sample. The effect size of their finding was very small (*d* = 0.15).

When analyses of sex differences are further refined to account for additional demographic variables such as participant age or socio-economic status, detailed results emerge with respect to verbal abilities. Recall that negligible differences emerged when Hyde and Linn (1988) assessed vocabulary across all ages of males and females. When they accounted for age and task type in their meta-analysis, a different picture emerged. In tasks of vocabulary, they identified that girls performed better than boys between six and 10 years of age (0.26*d)*, males and females performed equally well between the ages of 11 to 18 years, while between 19 to 25 years men seem to perform better than women (0.23*d*). This would suggest that age is a significant factor in the emergence of sex differences in certain types of intellectual abilities. This will be discussed further in relation to general intelligence and Lynn's developmental theory of sex differences (Lynn, 1999) in sections 2.3.2.5. and 7.1.1.

Another example illustrating how overall differences can mask task-specific differences is provided by Barnett & Rivers (2005). When they assessed verbal ability data of school children across the US in terms of overall differences, boys were found to underperform relative to girls. When the data was reanalysed separately by ethnicity and social class, a different picture emerged. They revealed that Caucasian and Asian boys did not differ significantly from girls. The Black and Hispanic boys were doing poorly, particularly those in underprivileged urban and rural areas, but relative to the other groups, so too were the girls of these groups. These results imply that sex may not be the causal element involved in the difference as commonly interpreted, but rather may be due to additional demographic variables such as, in this case, the underprivileged nature of the children's learning environment.

What is apparent from this brief review of the literature of spatial and verbal abilities is that claims of overall differences can be significantly affected by the measurement approach and specificity of analyses. By approaching the assessment of group differences in terms of an overall score may actually be obscuring nuanced differences at a more refined level, influenced by such variables as the task type, measurement method, sample composition and age of the sample. While a number of studies report a female advantage, the verbal abilities literature does not conclusively indicate that females are outperforming males on the same tasks, across age, all of the time. This brief review of the literature pertaining to verbal abilities highlights the need for further empirical studies that make use of large scale representative samples with clear measurement objectives and analyses that account for important demographic variables. Just as with other types of intelligence, there remains much to be understood about the nuances of the sex differences in mean performance of verbal and spatial abilities.

2.3.2. Sex Differences in General Intelligence

The general intelligence factor, *g*, is thought to be a common source of individual differences on all tests of ability. Despite its presence in all tests of ability, the assessment of sex differences in *g* is technically difficult to answer (Abad, Colom, Rebollo, & Escorial, 2004).

Just as with the literature on sex differences in spatial and verbal abilities, the question of sex differences in general intelligence continues to be an issue of considerable debate with respect to mean and variability of scores. Through a comprehensive review of the literature, it will become apparent that results are inconsistent, and appear to vary according to aspects of the research such as the method of assessment, sample composition, and age of sample. First a discussion of mean differences will be provided, followed by a discussion of variability.

2.3.2.1. An introduction to mean differences in g

As detailed in Chapter 1, general intelligence can be measured in a number of different ways by a number of different measures. According to the studies using these varied measures, some researchers conclude that there are no differences in the mean performance of males and females (Colom, Juan-Espinosa, Abad, & García, 2000; Jensen, 1998; van der Sluis et al., 2006).

For example, Rojahn & Naglieri, (2006) failed to find meaningful sex differences using the Naglieri Nonverbal Ability Test with a sample of children between the ages of six to 17 years. In another study of sex differences in *g* as measured by the Woodcock-Johnson III, Keith et al. (2008) found that inconsistent differences emerge across the age span of their sample (six to 59 years). They found an inconsistent difference for children (six to 11 years); for adolescents (12 to 17 years), they found a small non-significant difference in favour of females; and for adults (18 to 59 years) they found a significant female advantage.

One of the few studies in the literature that, prior to the comparison of mean differences, accounted for measurement invariance and ensured that there was no bias in the test towards either of the sexes was that conducted by van der Sluis et al., (2008). At the subtest level of the WISC, they determined that girls outperformed boys on Coding, while the boys outperformed girls on Information and Arithmetic. Despite these differences on the subtests, they concluded that there were no significant differences in *g*.

In contrast, other research claims a male advantage (Deary, Irwing, Der, & Bates, 2007; Lynn & Irwing, 2004). In their study of a very large validation sample of 17 to 18 year olds who took the Scholastic Assessment Test (SAT) in the US, Jackson & Rushton, (2006) concluded that, although the magnitude of difference in *g* between males and females was not large (d = 0.12), “it is real and non-trivial” (p.479).

These variable research findings are seen nowhere so strongly as in studies that employ the Raven’s Matrices. For many researchers interested in assessing sex differences in general intelligence, the Raven’s Matrices are often the measure of choice due to its reputation as one of the purest measures of general intelligence (Raven, 2009). In the extensive body of literature pertaining to the Raven’s all possible outcomes are evidenced: some conclude that males are the more intelligent sex overall, some find the females outperform males, while others maintain that there is no meaningful sex difference (Court, 1983; Mackintosh, 1996).

To date, much of the literature reporting significant sex differences made use of the advanced version of the Raven's Progressive Matrices (APM; Raven, Raven, & Court, 1998). The APM is intended for use with individuals of above average intelligence and is designed to reliably identify those individuals in the top 25% of the population. It consists of two item sets, where the problems have been arranged such that the items become progressively more difficult. Set I contains 12 items that are similar in difficulty to the SPM, and serve as a set of practice items for those completing Set II. Set II contains 36 items which are more difficult than those presented in Set I, and can be administered with or without a time limit.

While many would argue that the Raven's Matrices measure nothing but *g* (Jensen, 1998; Raven, 2009), others would dispute this (Carpenter, Just, & Shell, 1990; Lynn, Allik, & Irwing, 2004; Mackintosh & Bennett, 2005; van der Ven & Ellis, 2000), particularly with respect to the more difficult items of the APM. Using factor analytic evidence, it is contended that different cognitive components are employed to solve the items, with the easier items of the Matrices measuring a Gestalt or perceptual process while the more difficult items measure an analytic or analogical process. In the literature pertaining to sex differences in general intelligence, consideration is often given to the cognitive components involved in solving the test items and whether these differ by sex.

With respect to the APM, Carpenter, Just, and Schell (1990) evaluated which analytic processes differentiate between high and low scoring individuals, such as eye fixation patterns and errors they committed. They concluded that, common to all individuals in the sample, was an incremental strategy for encoding and inducing regularities in each test item. The processes involved in differentiating individuals of varying ability levels were the ability to induce abstract relations and to dynamically manage a number of different problem-solving goals in working memory. This resulted in the conceptualisation of the APM in terms of a two factor solution of the measure where items are solved using five different rules:

1. *Constant in a row*: the same value of an item attribute occurs throughout a row but changes between rows;
2. *Quantitative pairwise progression*: a quantitative change occurs in size, position or number of an attribute between adjacent

figures in a row;

3. *Figure addition or subtraction:* a figure from one column is added to, or subtracted from, another figure to produce a third;
4. *Distribution of three:* three values of a categorical attribute are distributed across the row; and
5. *Distribution of two:* two values of a categorical attribute are distributed across the row, with the third value being null.

DeShon, Chan, & Weissbein (1995) evaluated whether performance is dependent upon the same cognitive process for all of the items in the measure. They identified a number of rules employed in solving the test items that can largely be grouped into two different categories: Visuospatial and Verbal-Analytic.

There are six Visuospatial rules:

1. *Superimposition*: encoding and storing a perceptual representation of an object that is mapped onto a 2nd object by aligning the borders. The new image is comprised of the sum of the features of the two overlapping items.
2. *Superimposition with cancellation:* Much like the *Superimposition* rule, the new image is comprised of an object mapped onto a 2nd object by aligning the borders. Features that overlap cancel each other out. The final image is comprised of two overlapping object minus the features common to both.
3. *Object addition/subtraction:* the process of visually combining two objects into a whole. Objects are not superimposed, but place next to one another in accordance with a common border.
4. *Movement:* Objects move incrementally from frame to frame, giving the illusion of apparent movement within columns or across rows.
5. *Rotation:* The object must be mentally rotated in order for them to bring the object into correspondence across the rows or within columns. The same degree of rotation must be applied in order to solve the missing matrix.
6. *Mental transformation*: An operation is performed on an object

in the first entry that is specified in the second entry in order to yield the object in the final entry.

In addition to the Visuospatial rules, there are four Verbal-analytic rules: constant in a row, quantitative pair-wise progression, distribution of three, and distribution of two. These rules are operationally defined in the same way as defined by Carpenter et al., (as above).

Van der Ven and Ellis (2000) determined, by means of a Rasch analysis of the SPM that Set A and the first half of Set B measures a Gestalt perceptual process. They concluded that the second half of Set B and Sets C through E measure an analytic/analogical reasoning process. However, they concluded that it was the involvement of additional aspects that prevented them from confirming the Raven's as a unidimensional measure, namely in some of the items in sets C and E. In set C, they found evidence for "perceptual distraction", while in half of the items in Set E they indicate that a "coping strategy" was used by some of the participants. This coping strategy implies that the subjects make use of an easy, if incorrect, solution that doesn't take into account all of the item information available.

Another set of factor analyses of the SPM, conducted by Lynn, Allik, and Irwing (2004), confirmed the existence of a Gestalt perceptual factor in Set A and the first half of Set B, as found by van der Ven and Ellis (2000). However, they determined that the analytic/analogical factor found by van der Ven and Ellis could be sub-divided into two separate factors, in line with the factor solution suggested by DeShon, Chan, and Weissbein (1995) for the APM. The first, verbal-analytic reasoning is so named because the items involve aspects of addition and subtraction that requires verbal reasoning to be solved. The second, visuospatial ability factor contains items that require the item solutions to be found perceptually. With the exception of seven items from set A and one item from set B, all of the items loaded highly onto a second-order *g* factor.

Although the main focus of this dissertation remains the Standard Progressive Matrices, the literature pertaining to the Advanced Matrices is deemed relevant to the discussion of sex differences in general intelligence, and will therefore also be reviewed.

2.3.2.2. Male Mean Advantage

For many generations, a perspective has been maintained that males are the superior sex. This viewpoint has been particularly salient with respect to intelligence, and many researchers continue to uphold this belief. Results from a number of studies using the Raven's Matrices conclude a male advantage, and will now be described.

A meta-analysis conducted by Irwing and Lynn (2005) provides a comprehensive review of studies using the Raven's Matrices completed between 1972 and 2000. Of those reviewed, nine studies used the SPM (N_{SPM}=11,002) and 12 studies used the APM (N_{APM}= 9,196). The samples drew from 10 international locations and primarily comprised university students ranging in age from 18 to 46. They concluded that males outperformed females at the overall level (0.14*d*) as well as at the test level: for the SPM (0.10*d*) and for the APM (0.20*d*). From these findings, they disconfirmed claims others have made of 'no sex differences in general intelligence'.

However, the methodology employed in this meta-analysis has been raised into question (Blinkhorn, 2005), namely due to the exclusion of a particularly large sample from Mexico on the grounds of it being an 'outlier', and failing to weight the results of each study by their sample size. Blinkhorn further claimed that their use of the median of estimated differences multiplied by the general standard deviation as a "flawed and suspect tactic" (p. 32). He argued that a standard deviation of 15 is likely too large for a typical university sample, resulting in an inflated estimate of the male advantage. Irwing and Lynn (2006) addressed Blinkhorn's concerns, claiming that they found strong probability of bias due to moderator variables in the studies that they sampled. For example, they indicate that the Mexican study in question was largely male-biased and would likely underestimate the size of the population sex difference. The influence of the Mexican sample is evidenced by the considerable change in effect size when the sample is omitted (d = 0.31). Irwing and Lynn state this as the reason for using the median of the estimates in their calculations.

A further oddity in this set of analyses is the inclusion of part of their previously published meta-analysis (Lynn & Irwing, 2004). It appears as though they included four of the studies from the United States for further analysis in their more recent meta-analysis. In light of a number of analytical issues associated with conducting meta-analyses, one would question whether Lynn and Irwing's meta-analysis of meta-analytic data was a sound methodological approach to assessing sex differences in mean performance. The meta-analysis performed by Irwing and Lynn (2004) identified inconsistent differences across the age range of the participants, with a male advantage becoming more salient as age increases. Further details of this meta-analysis will be discussed in relation to the developmental influences of sex differences (section 2.3.2.5).

While there are many advantages to the process of systematically reviewing a number of related studies in a meta-analysis, it does also have a number of weaknesses, namely the problems of "Apples and Oranges", "Garbage in, Garbage out", and "the File Drawer" (Hyde & Linn, 1988; Sharpe, 1997). The "Apples and Oranges" problem, as it is known, refers to the fact that a number of disparate research issues are aggregated and averaged, as can be seen in meta-analyses of multiple cognitive abilities (e.g., Hyde & Linn, 1988). In order to address this problem, it is best to conduct meta-analyses on a narrow range of abilities or latent constructs.

Another problem with meta-analytic methods is the "Garbage in, Garbage out" problem. This problem refers to the inclusion of poor quality studies in meta-analyses, such as those samples that are not considered representative of the general population or are of a particularly small sample size. The inclusion of poor quality studies may interfere with the ability to establish and correctly identify relationships that exist within the studies of good quality.

A third problem associated with meta-analytic studies is the "File Drawer" problem. This issue relates to the fact that studies showing significant effects tend to be published, while studies showing non-significant effects tend not to find their way to the literature. In light of meta-analytic studies, this is a particular concern due to the fact that the resulting analyses will largely be unbalanced by an under-representation of studies showing non-significant effects. This will be discussed further in section 2.3.2.4 in relation to the lack of studies in the literature that showing no significant sex differences in mean performance in general intelligence.

While the overall conclusion drawn by Lynn and Irwing (2005) is that males outperform females, this is not to say that each individual study shows a male advantage. Further, in light of the critique of the meta-analysis they conducted, it is worthwhile reviewing the finer points of some of the sex differences literature including a number of the studies they reviewed.

Mohan and Kumar (1979) used the SPM in their assessment of the role of neuroticism in learning and performance. In a sample of 400 participants ($N_{neurotic}$ = 200, $N_{non\text{-}neurotic}$ = 200) they indicate that the non-neurotic sample perform significantly better than do the neurotic participants, particularly with respect to the more difficult items on the test. They then go on to present results of the males and females of the sample (N_{male} = 200, N_{female} = 200). There were no significant differences in performance on the SPM between the males and females of this sample (p = 0.43). Although they themselves do not make any claims about differences in performance between the sexes, it raises the question of whether a study such as this ought to be included in meta-analyses with the overall aim of determining sex-differences in intelligence.

One of the main findings of this study indicates that neurotic participants do not perform as well as non-neurotic participants. In the literature, women are noted as scoring higher on neuroticism than males (Fanous, Gardner, Prescott, Cancro, & Kendler, 2002). It therefore follows that females, by nature of the greater neurotic tendency, may be disadvantaged in this study. It could therefore be argued that is not appropriate to include such a study in a meta-analysis of general intelligence.

Silverman et al., (2000) used an abbreviated 26-item version of the SPM as part of a study investigating mechanisms underlying way-finding in spatial ability. From their sample of 111 university students (65 females and 45 males, mean age = 22 years), they concluded “unexpectedly” (p. 210) that males outperformed females ($Mean_{males}$= 16.57; $Mean_{females}$= 14.77; d = 0.87). Due to the unrepresentative nature of their sample, these results must be interpreted with caution.

Similar evidence can be found in the literature pertaining to studies using the APM, where the large majority of studies report a male advantage in mean performance. In the process of developing a short-form of the APM suitable for university undergraduates, Bors and Stokes (1998) administered

Sets I and II under timed conditions to 506 undergraduate students (N_{male} = 180, N_{female} = 326) at the University of Toronto. The performance on Set I was found to be negatively skewed, leading them to conclude that the item set was relatively easy for the participants of their sample. They found this surprising in light of the fact that the items are intended for use across ability levels. In terms of mean performance, they determined that males slightly outperformed females, but not significantly so ($d = 0.12$; $p > 0.05$). When the sample was administered Set II, the males were found to significantly outperform the females ($Mean_{male} = 23.00$; $Mean_{female} = 21.68$, $p < 0.05$).

Although not the primary research aim for some, investigations of sex differences are conducted as secondary research aims. While the primary aim of the investigation by Rushton and Skuy (2000) was to assess performance differences between African and White university students in South Africa, they discovered a small male advantage in both African and White samples using the SPM. Their sample consisted of 309 17- to 23-year-old university students. Overall, males obtained a mean score of 50.3, while females obtained a mean score of 47.7 out of 60 items[6]. Further analysis of their results (Mackintosh & Bennett, 2005) revealed that the noted sex differences were only attributable to performance differences on a small number of items.

While the socio-economic backgrounds of the different groups of the sample are noted in their paper, Rushton and Skuy did not account for them in their analyses. Performance on tasks of a visual-spatial nature has been linked to social, cultural, and environmental factors (Bradley & Corwyn, 2002; Johnson, McGue, & Iacono, 2007), while indices of economic development and gender equality having been shown to be strongly correlated with mean performance on tasks of mental rotation. Although a male advantage in mental rotation is evident cross-culturally, groups demonstrating greater gender equality and economic development were associated with better mean performance (Lippa, Collaer, & Peters, 2010). Such findings could be implied in the findings of Rushton and Skuy (2000).

[6] Details provided in the article were such that calculation of effect size of the difference between males and females was not possible. However, for Black male and females, $d = 0.51$, and for White males and females $d = 1.62$.

While much of the Raven's literature approaches analysis from the perspective of a unidimensional construct (e.g., Lynn, 2004; Silverman et al., 2000; Mohan & Kumar, 1979), others would contend that the Matrices are multi-dimensional (as described in section 2.3.2.1), measuring such constructs as visuospatial ability, and verbal-analytic strategies. Throughout the literature of sex differences in spatial abilities, the general consensus is that males outperform females (Hyde, Fennema, & Lamon, 1990; Maccoby & Jacklin, 1974; Voyer, Voyer, & Bryden, 1995). Some researchers (such as DeShon, Chan, & Weissbein, 1995; Lim, 1994; Schweizer, Goldhammer, Rauch, & Moosbrugger, 2007) believe that it is the visuospatial quality of some of the items on the Raven's Matrices that account for the existence of a sex difference on the measure.

To assess whether the sex differences on the Progressive Matrices were influenced by sex differences in spatial ability, Colom, Escorial, and Rebollo (2004) administered the APM to 239 undergraduate university students along with the Spatial Rotations Test. They concluded that males outperformed females on both measures: $d_{\text{APM}} = 0.29$; $d_{\text{SRT}} = 0.57$. However, when the sex difference in spatial ability was controlled for by analysis of covariance, the difference in performance of males and females on the APM was non-significant ($p = 0.393$). It has been suggested that, due to the visuospatial nature of some of the items of the Matrices, these items can be considered biased against females.

In order to address the question of visuospatial bias in the Matrices, Abad, Colom, Rebollo, and Escorial (2004) administered Set II of the APM under a time limit to 1970 university applicants. It is not clear from the sample description which subject(s) or programme(s) the participants were applying to. They first aimed to determine whether a one or two factor solution to the model was appropriate for their data. They compared a one-factor model with two different two-factor models (Deshon, Chan, & Weissbein, 1995; Dillon, Pohlmann, & Lohman, 1981). They concluded that the multi-factor models did not fit the overall data set better than the one-factor solution, and therefore, proceeded with an assessment of Differential Item Functioning (DIF) using a unidimensional model. To account for the different item types, they sequentially analysed different item groupings. They determined that of the four items identified by Deshon, Chan, &

Weissbein (1995) as verbal items, two of these were found to be easier for females and biased against males. When they assessed the visuospatial items, they determined that 45% of them had non-uniform DIF and were biased against females. Upon reanalysis after the removal of the biased items, a significant male advantage was still present. Although the ultimate conclusion from this research is that the male advantage on performance of the APM is not attributable to the biased items, they hypothesised that the unbiased items might still be measuring visuospatial ability to some extent. Because the nature of the sample that they used could have been influencing the outcome of their analyses, they recommended that analyses be replicated using a more representative sample.

A further study questioning what the Raven's Matrices measure was conducted by Mackintosh and Bennett (2005). They analysed the results of 97 secondary school students of 17 to 18 years of age who completed Set II of the APM, the Differential Aptitude Test (DAT) verbal reasoning measure, and the Mental Rotations Test (MRT). In terms of the overall performance, they found that males significantly outperformed females on the Raven's (*d* = 0.43, p < 0.05). What they also found was that this was true only on certain types of Raven's items. Ascribing to the item solving rules put forth by Carpenter, Just, and Schell (1990) (which are also quite close conceptually to those rules proposed by DeShon et al., 1995), they re-assessed the Raven's data according the type of answering strategy required to solve the items: Pair-wise Progression, Distribution of Three, Addition/Subtraction, and Distribution of Two. They determined that males outperformed females on all four item types, with the largest significant difference in performance occurring on items requiring the 'Distribution of two' strategy (*d* = 0.45, p < 0.05), followed by the 'Addition/Subtraction' items (*d* = 0.60, p < 0.01)[7]. On the items that used 'Distribution of Three' (*d* = 0.10, p = 0.51) and 'Pair-wise Progression' (*d* = 0.17, p = 0.17), the performance difference was small and non-significant. When they associated the performance on the Raven's with the performance on the

[7] Although the effect size of the difference in the Addition/Subtraction items is numerically larger than the Distribution of Two items, this was attributed to the greater variance of the Distribution of Two items.

DAT and the MRT, interestingly, they found that males' performance was more highly related with the MRT than their DAT scores. The opposite was true for women. These findings correspond to the literature pertaining to sex differences in visuospatial ability (e.g., Voyer, Voyer & Bryden, 1995) and verbal abilities (e.g., Hyde & Linn, 1988).

Results such as these have led to suggestions that males and females employ different strategies to the solving of the Raven's Matrices items. For example, Lim (1994) determined that the factor structure of the Raven's matrices varies according to gender, resulting in differential performance. Despite this, Mackintosh and Bennett refute any suggestion that their correlational results suggest that the Distribution of Two or Addition/Subtraction items contain a stronger spatial component than other items on the test.

A number of different factor structures, or taxonomies, have been proposed in the literature pertaining to the different underlying strategies used for solving different items of the Raven's Matrices. These were described in an earlier section of this text. The most commonly cited taxonomies in the literature are by DeShon, Chan, and Weissbein (1995), Dillon, Pohlmann, and Lohman (1981), and Carpenter, Just, and Shell (1990), which was also tested by Mackintosh and Bennett (2005). The pattern of sex difference found by Mackintosh and Bennett (2005) is considered to be similar to that found by DeShon et al. (1995) with their distinction between visuospatial and verbal-analytic items.

Vigneau and Bors (2008) tested the fit of a number of different factor models, including a unidimensional model, in order to verify the previous findings of Mackintosh and Bennett (2005) and of DeShon et al. (1995). An opportunity sample of 506 first-year psychology undergraduate students (N_{male} = 180, N_{female} = 326; $Mean_{age}$ = 20.00 years) were assessed with Set II of the APM under timed conditions. Their Confirmatory Factor Analyses concluded that the unidimensional model provided the best fit of their data, and therefore failed to find support of the multi-factor taxonomies proposed earlier in the literature. The one-factor solution concluded that males significantly outperformed females (d = 0.24; $p < 0.05$). When they proceeded to assess the data according to the taxonomy proposed by Mackintosh and Bennett (2005), they found significant differences between the sexes for Pair-wise Progression items (*d* = 0.20; p < 0.05) and the Distribution of Two items (*d* = 0.32; p < 0.05). For the Addition/Subtraction and the Distribution of Three items, the males outperformed the females, but not significantly so ($d_{A/S}$ = 0.04; p > 0.05; d_{D3} = 0.16; p > 0.05). The replication of the DeShon et al. (1995) taxonomy led them to conclude that males significantly outperformed females on Visuospatial items (*d* = 0.26; p < 0.05) but not on Verbal-analytic items (*d* = 0.16; p > 0.05). However, when they attempted to reanalyse the data in terms of the association of gender and item score, they noted that the gender differences in their data were found to be dispersed across the items rather than forming groupings as suggested in the literature.

2.3.2.3. Female Mean Advantage

In contrast to the commonly reported finding in the literature that males are outperform females in measures of general intelligence, the study conducted by Abdel-Khalek and Lynn (2006) of eight to 15 year-olds in Kuwait really stands out. It is, to date, the only published study of the SPM that shows a female advantage across all age groups of their sample (eight to 15 years). Although the difference they found was small (*d* = -0.08), it is highly significant (*p* < 0.001) and the sample size was large (N = 6529; N_{male} = 3278, N_{female} = 3251). These results are further supported by their finding that the females outperformed the males on additional measures of verbal comprehension, foreign language ability, and mathematics.

A further study reporting a female advantage is available from a large representative sample of six to 11 year old children in the United Arab Emirates assessed with the Coloured Progressive Matrices (CPM; Khaleefa & Lynn, 2008). Recall that the CPM is a simplified version of the Standard Matrices intended for use with children or individuals of impaired intellectual functioning. Results from the very large sample (N = 4,496) suggest that females outperform males in the 12 age groups tested by approximately two IQ points. However, the analytical strategy and score conversion practices employed by Khaleefa and Lynn is somewhat unconventional in the literature.

Rather than making use of the published CPM standardisation data (Raven, Raven, & Court, 1998b), they made use of the British SPM normative data from 1979 and claimed to do so because they “are more accurate than the Coloured Progressive Matrices” (p.59). Further, as the Standard Matrices are intended for individuals from seven to 18 years of age, normative data for the SPM was not available for the six year olds in their sample, so the 1982 British normative data of the CPM was used for the six year olds. While numerically they provide indication that females outperformed males in their sample, the analytical methods that they employed do not provide sound statistical grounds upon which an argument for a female advantage can be confidently made.

These two studies do not provide sufficient evidence of a female advantage to counter the large one-sided literature showing a male advantage. However, what it does provide is an indication that there is more to be understood regarding sex differences in general intelligence. More robust, psychometrically sound studies are needed in order to gain a better understanding of the conflicting empirical results evidenced in the literature.

2.3.2.4. No Mean Difference

In contrast to the findings that show a male advantage or a female advantage, there are studies in the literature that indicate that no significant differences exist on the SPM between males and females. However, in comparison to the abundance of studies in the literature claiming a male advantage there are relatively few published studies of the SPM that show no meaningful differences.

A recent study by Rushton and Cvorovic (2009) reported no significant sex differences on the Raven’s standard matrices on four samples of adults in Serbia (17 to 65 years of age). Interestingly, their sample comprised 418 males and 190 females (N=608). There was, therefore, an uneven distribution of males to females in the sample, which may have affected the overall outcome of the results. It would be particularly interesting to see what the results would have been had a matched sample of males and females been used in the analyses. Further, Rushton and Cvorovic compared the results of their sample to the 1993 US edition of the measure. They concluded that the Serbians were underperforming relative to an American standardisation sample ($IQ_{Serbian}$ = 90 versus $IQ_{American}$ = 100). This finding raises the ethical issue of using tests cross-culturally.

Sound psychometric testing practices advocate that tests ought to be used only with the populations for which they were designed or standardised with. This again reinforces the issue that such results may not be optimally generalisable.

Crucian and Berenbaum (1998) conducted an investigation of sex differences in right hemisphere tasks that included the SPM in their test battery. In their sample of 86 males and 132 females, they failed to find a significant difference between the sexes on performance on a modified version of the SPM[8] *(d* = 0.24). As there was no information provided on the modifications they made to the SPM, these results should be considered with caution.

A standardisation of the SPM was conducted recently in Syria by Khaleefa and Lynn (2008b). In the sample of 3,489 individuals between seven and 18 years of age, they conclude no significant sex differences (*d* = -0.06). They did, however, find a significant male advantage at 11 years of age (*d* = 0.47) and a female advantage at 16 years of age (*d* = -0.45). They attribute these findings to be the result of sampling error.

Data for 14 year old females was not available, yet the 136 participants in this category were included in the total reported number of participants. Reasons for the lack of data were not presented. They did, however, include the data from the 14 year old males, resulting in unequal proportions of males and females in their total data set: N_{female} = 1613, N_{male} = 1739. The exclusion of female data for this age group could have significantly influenced the overall findings of this study, which should therefore be interpreted with caution.

A study by Lynn, Backhoff, and Contreras-Niño (2004) also failed to find significant differences in the SPM as a whole in a sample of 920 seven to 10 year old children in Mexico (*d* = 0.09). They identified a trend in the scores, whereby the slight male advantage that was apparent at seven years decreased to a point of non-significant female advantage at 10 years. The developmental involvement in sex differences in intelligence will be

[8] Authors do not specify they ways in which the SPM was modified, but do indicate the version used in the study is reliable (Crucian & Berenbaum, 1998).

discussed further in section 2.3.2.5. However, it is possible to question the quality of the analyses they conducted, as the basis of their conclusions are upon analysis of variance, which are now acknowledged to be lacking the strength and specificity to correctly identify sex differences (Embretson & Reise, 2000).

A "file-drawer" explanation has been offered as the potential reason for the lack of studies reporting no sex differences on the SPM. It has been suggested that studies finding no sex difference in intelligence are deemed note as interesting as significant findings, and therefore not published as frequently (Blinkhorn, 2005). It has even been suggested that the only scientists publishing on the topic of sex differences in intelligence are "those whose agenda is to prove women … intellectually inferior" (Begley, 2009; p. 53). If true, this would result in the over-reporting of studies whose results show a male advantage on measures of general intelligence, which has clearly been evidenced in the current review of the literature. The prevalence of such findings in the literature could arguably contribute to societal perceptions of ability and, ultimately, the gender discrepancy evident the wage gap and differences in career advancement as noted in section 2.2.

In reviewing the results of these studies, it is particularly important to be mindful of the sample characteristics, namely age and representativeness of the participants. What is salient about a number of these studies is that they employ convenience samples (also referred to as opportunity samples or accessible populations) – samples of participants that are readily available and convenient rather than representative of the normal population. For example, Rushton and Skuy (2000) rest their conclusions upon an opportunity sample of 309 university students in South Africa.

Another example discussed in section 2.3.1.2, Weiss, Kemmler, Deisenhammer, Fleischhacker, and Delazer (2003) based their conclusions of an overall female advantage in lexical fluency upon a sample of 97 students studying psychology and medicine at University. However, it is unlikely that the students in their study are typical of the average university student or of an individual in early adulthood[9]. While such samples have

[9] The age of the participants was not provided in the article but was assumed according to the typical

their use (such as for pilot testing new measures), they generally result in a largely homogeneous sample.

It is fair to say that these types of samples are non-optimal for making generalisations to the population as a whole for they lack representativeness of the wider population (Hunt & Madhyastha, 2008), particularly in light of the fact that the implication of sex differences in intelligence can have potentially widespread societal ramifications. Further, it is probably that a ‘file-drawer’ problem exists in the literature of the SPM, resulting in a largely one-sided literature based upon predominantly opportunity samples.

The external validity of research conclusions based upon opportunity samples is largely unknown (Fan, Chen, & Matsumoto, 1997). Consideration of such convenience samples highlights the need for investigations of sex differences in general intelligence that make use of samples representative of the population as a whole in order for the results to be optimally meaningful. In light of these facts, there is clearly much to be understood about sex differences in general intelligence and on the Raven’s Standard Progressive Matrices.

2.3.2.5. Sex differences and the Developmental Trajectory

Another way to approach investigations of sex differences than the traditional search for an overall male or female advantage is within the context of a developmental trajectory: boys and girls mature at different rates, and differences in intelligence may be influenced by the variation in development. For generations, it has been suggested that girls mature earlier than boys with respect to the development of a number of physical and cognitive characteristics (Hohm, Jennen-Steinmetz, Schmidt, & Laucht, 2007; Nature, 1923; de Onis et al., 2007). According to Lynn (1999), failure to account for differences in maturation between boys and girls may be masking true sex differences in general intelligence in a number of studies of children and young adults.

age of an undergraduate student.

According to the maturational differences, Lynn hypothesised that a female advantage would be evident pre-puberty, while a male advantage would begin to emerge after puberty in late adolescence or early adulthood. His “Developmental Theory of Sex Differences” (Lynn, 1999) proposed that girls mature earlier than boys, both cognitively and physically, and tend to have a cognitive advantage over males of about 1 IQ point between eight to 15 years. By 15 years of age, however, there is a developmental deceleration for females while boys continue to develop. Lynn claims that this results in a male advantage of approximately 2.4 IQ points from approximately 16 years of age, an advantage that is maintained throughout adulthood (Lynn, 2002; Lynn, Allik, & Must, 2000).

A number of studies of the SPM do appear to provide evidence that suggests a developmental trend, with intellectual advantages emerging at different ages for boys and girls. However, the results do not necessarily show the same developmental pattern or direction of difference as clearly as delineated by Lynn. This has resulted in a collection of studies that tends to provide contradictory findings.

In a sample of 12 to 18 year olds in Estonia, Lynn, Allik, and Irwing (2004) determined that girls performed better than boys from 12 to 13 years of age (*d* = -0.384)[10], there were no differences between the sexes between 14 to 16 years of age (*d* = -0.033), but a male advantage emerged at 17 years of age (*d* = 0.193). Using a standardisation sample of the SPM in Estonia, (Lynn, Allik, Pullmann, & Laidra, 2004) conclude a female advantage among 12 to 15 year olds (*d* = -0.03 to -0.54), and a male advantage between 16 to 18 year olds (*d* = 0.04 to 0.80).

Abdel-Khalek and Lynn (2006) conducted a study using a large SPM standardisation sample of eight to 15 year olds in Kuwait. They determined that girls outperformed boys between eight to 12 years (*d* = -0.06 to -0.27) when a small non-significant male advantage between 13 to 15 years (*d* = 0.01 to 0.06) began to emerge. A further study by Lynn, Backhoff and Contreras-Niño (2004) identified a developmental trend in the scores of a

[10] A positive effect size (+) represents a male advantage while a negative effect size (-) is indicative of a female advantage.

large sample of seven to 10 year old children in Mexico. A slight male advantage was apparent at seven years but this decreased to a point of a non-significant female advantage at 10 years.

Evidence of a developmental theory is also available from meta-analytic reviews of the literature. Lynn and Irwing (2004) conducted a meta-analysis[11] to summarise a number of different studies of the SPM, some of which have already been discussed (e.g., Mohan & Kumar, 1979; Crucian & Berenbaum, 1998). They concluded that, across the studies, boys obtain slightly higher means than girls from six to nine years of age, but not significantly so (*d* = 0.01 to 0.10). From 10 to 13 years, a higher non-significant mean emerges for females (*d* = -0.06 to 0.05). At 14 years of age, a male advantage emerges (*d* = 0.08) which, at 15 years, becomes significant and increases in effect size to 0.10*d*. By 18 years of age, the significant difference increases in size to 0.16*d*.

Within the field of sex differences in general intelligence, the subject of a developmental theory is a relatively novel concept. The only published studies using the SPM demonstrating a developmental trend have been conducted by Lynn and his colleagues, which do not offer a particularly unbiased body of evidence. Evidence of developmental trends in sex differences in general intelligence using other measures is available, but only one of these studies did not involve Lynn.

Using the Naglieri Nonverbal Ability Test (NNAT), Rojahn and Naglieri (2006) concluded findings that were consistent with Lynn's theory of developmental sex differences: there was no sex difference in their sample between six and nine years of age; they found a slight female advantage between 10 and 13 years; and between 15 and 16 years there was a male advantage. However, the effect sizes were much smaller than those found by Lynn, leading them to deem them of no practical importance. Their ultimate conclusion was that there was no meaningful sex difference in general intelligence at any age between six and 17 years.

Using a large standardisation sample of the Differential Aptitude Test

[11] The Lynn and Irwing (2004) meta-analysis also included studies of the Coloured Matrices and Advanced Matrices. A discussion of these findings was not deemed relevant.

(DAT) in Spain, Colom and Lynn (2004) provide evidence to support Lynn's developmental theory. While the overall advantage was attributed to males (4.3 IQ points), there was a significant female advantage between 12 to 14 years of age. This advantage diminished relative to boys from approximately 15 to 18 years.

Using a historical data sample of nine to 15 year old children tested with an Estonian adaptation of the American National Intelligence Test, Lynn, Allik, and Must (2000) argue for a developmental trend of intelligence. They conclude that at nine years of age, there is a small and non-significant sex difference (*d* = 0.22). At 10 years of age, there is a significant female advantage (*d* = 0.32), followed by a male advantage from 11 to 15 years (*d* = 0.09 to 0.37). They attribute these differences intellectual functioning to developmental sex differences in brain size and stature, which decrease between seven to 14 years, but increase from 15 to 18 years.

Unlike the theory proposed by Lynn, it is evident from the studies discussed above that there is considerable variation in the age at which sex differences emerge for males and females. One study of the SPM found evidence of a female advantage from 12 years to 13 years, while another study from the same country found this to be true of 12 to 15 year olds. Others found that girls outperformed boys from eight to 12 but that no differences existed between 13 to 15 years. Even amongst the studies making use of measures other than the SPM, there lacks consensus about the ages at which consistent sex-linked differences in intelligence emerge.

Some would argue that due to the fluctuation in the magnitude of sex differences at different ages, the small effect sizes and inconsistent findings provides evidence against the hypothesis for sex differences in general intelligence (Hyde, 2005). Others would claim that such findings are conclusive evidence of sex differences. Others still would argue that such findings are evidence of the need for more understanding of the nuances of sex differences. As the Raven's Matrices are disputed as one of the purest measures of general intelligence (Jensen, 1998), there remains a need in the literature to further verify the claims of a developmental theory of sex differences using the SPM with a representative generalisable sample. This dissertation addresses this need and it will be assessed further in Chapter 7.

2.3.3. Score Variability in Intelligence

Another issue to consider when discussing sex differences in cognitive ability is the variability of scores. The vast majority of studies to date have focused upon sex differences in mean performance while implicitly assuming homogeneity of variance. This is evidenced in those studies making use of classical statistical methods and inferential statistics (such as *t* ratios; Feingold, 1992). This is an erroneous assumption which can have considerable influence on the overall conclusions drawn.

When score variability is considered in the literature, many would contend that the sexes differ in terms of the distribution of scores if not in mean performance. If there is a difference between males and females in terms of score variability on measures of intelligence then the more variable group is likely to be over-represented in both the low and high score levels, even if the mean score of both groups is the same (Feingold, 1992). The vast majority of studies maintain that males are more variable than females, with a greater proportion of males in the top and bottom ends of the overall normal distribution of scores (Arden & Plomin, 2006; Johnson, Carothers, & Deary, 2008; Eysenck, 1981). This is especially true of studies using measures of mathematics and spatial ability.

A landmark contribution to the study of sex differences in score variability was a review conducted by Hedges and Nowell (1995). They conducted secondary analysis of six studies[12] using large representative samples. They concluded that, while mean differences on a variety of cognitive abilities were small, the scores of males were consistently more variable. Amongst high scoring individuals, males tended to outnumber females on most tasks except in the domains of reading comprehension, perceptual speed, and associative memory (Hedges & Nowell, 1995). The implication of such findings suggests that if differences in intelligence were held constant between males and females, there would be more males than females with IQ scores below 70 and greater than 140.

[12] The Project Talent data set, the National Longitudinal Study of the High School Class of 1972 (NLS-72), the National Longitudinal Study of Youth (NLSY), the High School and Beyond 1980 data set (HS&B), the National Educational Longitudinal Study of the 8th Grade Class of 1988 (NELS:88), and the trend data sets for the National Assessment of Educational Progress (NAEP).

Similar findings were concluded by another landmark study of score variability conducted by Feingold (1992). He assessed the variance ratios of five different national test standardisation samples[13]. Each of the tests measured different aspects of intellectual functioning such as English proficiency, perceptual speed, abstract reasoning, mathematical ability, and non-verbal abilities. However, unlike many other studies in the literature, Feingold accounted for sex differences in both mean performance and variance simultaneously in order to obtain the most accurate reflection of the performance differences across the distribution of scores.

Across the measures, Feingold concluded that on tasks of quantitative reasoning, spatial visualisation, spelling, and general knowledge males were found to be consistently more variable than females. On tasks of verbal ability, short-term memory, abstract reasoning, and perceptual speed, there was found to homogeneity of variance.

When the sex differences in variance and mean performance are taken into consideration together, sex differences in the left tail and the right tail of the distributions must be evaluated separately in order to get an accurate account of the differences. Among the low-scoring proportion of the sample, the most notable difference was a large male advantage in mechanical reasoning. In the right tail of the distribution, among those with the highest scores, the notable sex differences were a male advantage in mechanical reasoning, and information, as well as a female advantage in perceptual speed and spelling.

Like Feingold, Nowell & Hedges (1998) conducted an assessment of sex differences that accounted for the mean performance and variance simultaneously in order to obtain a better understanding of the extremes of the distribution of scorers. They reviewed eight representative samples of twelfth grade pupils in the United States, with a view to better understand whether sex differences were changing over time (between 1960 and 1994). In doing so, they concluded that the overall gender difference in mean performance across measures was small (less than d = 0.3). The pattern of

[13] The Differential Aptitude Tests (DAT), the Preliminary Scholastic Aptitude Test (PSAT), the Scholastic Aptitude Test (SAT), the Wechsler Adult Intelligence Scales (WAIS, WAIS-R), and the California Achievement Test (CAT).

performance differences confirmed patterns as reported in previous studies in the literature (see section 2.3.1. of sex differences in different cognitive abilities). Females were found to outperform males on measures of reading, perceptual speed, and writing, while males outperformed females on mathematics, and science.

With respect to variance, their review concluded that males were more that 50% more variable than females on nearly all measures except for vocabulary. In terms of the extreme portions of the samples assessed, males were over-represented in the upper tail of the distribution in terms of mathematics and science, while females were over-represented in terms of reading, vocabulary and perceptual speed score distributions. In the lower tail of the distribution, males were over-represented in reading, perceptual speed, and writing, while females were over-represented in mathematics and science score distributions. Their overall evaluation was that the sex differences in mean and score variability were very small and consistent across time.

Unlike the number of studies of sex differences in variability of multiple cognitive abilities, there are relatively few studies that investigate such differences in general intelligence. Further, many reviewers have contended that investigations of sex differences of cognitive ability ought to be conducted on samples that are representative of the population due to the fact that opportunity samples might be confounding the results. One such study by Deary, Thorpe, Wilson, Starr, and Whalley (2003) addressed both of these by conducting an assessment of sex differences in mean and variability on virtually all children born in Scotland in 1921 and tested in 1932 (at 11 years of age) using a measure of general intelligence.

The Scottish Mental Survey 1932 (SMS32) comprised three different tests (two pictorial assessments, and one general ability measure) and was administered to 87,498 children. Despite the very large sample size of the groups of males and females, no difference in mean performance was evident. However, the distribution of scores was significantly different between males and females, with a slight over-representation of females who obtained IQ scores of 90-115. In the very low (IQ 50 to <60) and very high achievers (130 to <140) on the test, there were significantly more boys than girls with a ratio of approximately 1.4 boys per girl. While these research conclusions are relevant to the specific cohort for whom it was collected, such findings are consistent with other current reports of sex differences in mean performance and variability. Along with the findings of Nowell and Hedges (1998), these findings suggest that sex differences in variability are not a modern phenomenon but are historically evident.

The study by Deary, Thorpe, Wilson, Starr, and Whalley (2003) was further extended by Johnson, Carothers, and Deary (2008) who reanalysed the SMS32 data along with the SMS47 data collected from the 1947 cohort of 11-year-old children in Scotland. They first noted that the data of both the SMS32 and SMS47 were not normally distributed, each having negative skew (-0.211 and -0.347 respectively). In light of the very large sample sizes, this level of skewness is considered highly significant. The data was also found to be platykurtic, with small tails at the ends of the distribution, with kurtosis values of -0.581 and -0.495 respectively. In concert with the negative skew of the distribution, Johnson et al. (2008) conceptualise the data in terms of a mixture of two normal distributions: one that reflects those who are of impaired general intellectual functioning, and one that reflects those of normal intellectual functioning.

The results showed greater variability in the lower ends of the distribution of overall general intelligence among males compared to females, leading them to conclude that males were over-represented in the distributions representative of individuals with disrupted intellectual functioning compared to females. They also concluded that there were higher proportions of males at the higher ends of the distribution of general intelligence.

Studies of sex differences in variability of general intelligence using the Raven’s Matrices are very few. The limited evidence that is available from the SPM suggests that, if either sex shows greater variability, it is at least as likely to be females as males. In their meta-analysis of 9 studies that detailed the variance on the SPM, Irwing and Lynn (2005) found that females were more variable than males in 7 studies, and that there was no sex difference in variability in 2 studies.

Arden and Plomin (2006) conducted a study of mean and variability using data of children between two and 10 years of age from the Twins Early Development Study (TEDS). A number of different measures were used, including the SPM. However, the SPM was only administered to 10 year olds. At this age, analyses revealed that boys have a slightly higher mean than girls at 10 years of age (d = 0.10) and have slightly higher variability than girls. At the extremes of the distribution, the boys were only found to be over-represented in the top 10%. They remained in approximately equal proportion with girls in the lower 10% of scores.

As most of the literature has been devoted to the understanding of mean differences, there is relatively little known about the differences in the variability of scores between males and females (Brody, 1992). Analyses that provide a clear understanding of the distribution of scores in samples representative of the population are rare (Johnson, Carothers, & Deary, 2008). Yet, the issue of score variability may prove to have considerable social consequences. For example, educational institutions may make use of high test cut-off scores as part of their admission selection process. Depending on the test and the variability of scores, more males may meet the minimum score criteria than females, thereby rendering the admissions process discriminating against females (Brody, 1992). This might then result in an over-representation of males in certain educational programmes, and may be implicated as a reason for the noted gender imbalance in certain academic and professional careers, as well as upper-levels of employment (Fan et al., 1997). Such gender discrepancies could prove to be associated with further gender-related discrepancies discussed in section 2.2 such as the wage gap that exists between males and females who are employed in equivalent positions. Consequently, a greater understanding of the finer details associated with the ways that males and females are evaluated in terms of cognitive ability and selection processes could contribute towards a means for equalising the existing biases and discrepancies that exist in society today. This identified need in the literature will be address in this dissertation, with a review of score variability for the overall sample (section 6.2.1), and with respect to age (section 7.2).

2.4. Explanations for sex differences

Sex differences in cognitive ability, and the inconsistent findings, have been demonstrated extensively throughout the literature. Many explanations have been offered to explain these differences from the perspectives of biology, environment and life experiences, as well as combination of these.

Biological explanations attribute the differences to genetic determinants (Docherty, Kovas, Petrill, & Plomin, 2010; Johnson, Carothers, & Deary, 2008; 2009; Petrill, 1997) and hormonal influence prenatally and throughout the individual's life (Hines, 2010; Kimura & Hampson, 1994). Anatomical and structural differences in the brain have also been controversially implicated including differing brain sizes between the sexes (Cahill, 2006; Lynn, 1994; 1999; Lynn, Allik, & Must, 2000).

Different life experiences and societal gender stereotypes are often implicated in the debate surrounding cognitive sex differences (Ceci, Williams, & Barnett, 2009; Désert, Préaux, & Jund, 2009). The biopsychosocial perspective (Figure 3) argues for a collectivist approach that accounts for both the nature of biology and genetics and the nurture of society and life experiences in its explanations of the differences between boys and girls.

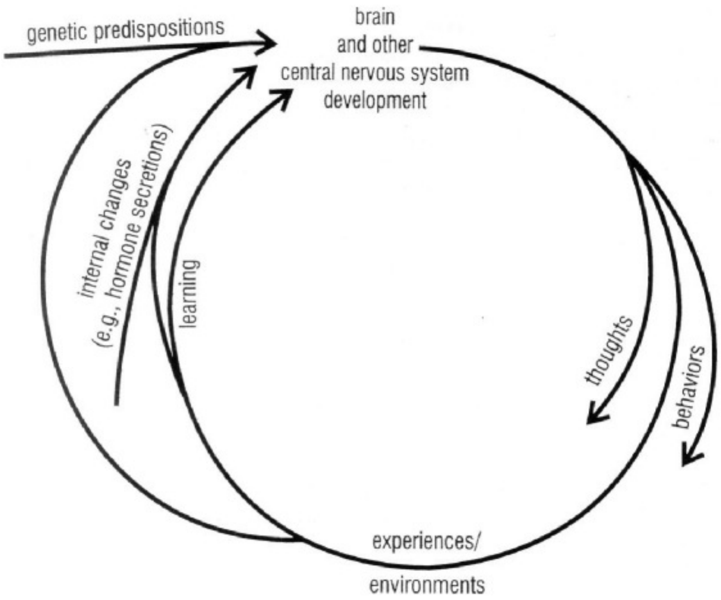

genetic predispositions
brain and other central nervous system development
internal changes (e.g., hormone secretions)
learning
thoughts
behaviors
experiences/ environments

Figure 3: A biopsychosocial model as a framework for understanding cognitive sex differences (Halpern, 2000; 2004)**.**

While there remains to be a definitive conclusion improvements in understanding hormone and gene function as well as advances in functional magnetic resonance imaging suggests that there may be substantial developments in the near future to the understanding of the origins of sex differences in intelligence. A further potential explanation of sex differences in cognition most related to this dissertation is a psychometric one. It suggests that some of the sex differences in cognition noted in the literature may be explained as artefacts associated with the way the test data is collected or how it is statistically analysed.

When data samples are not representative of males and females in general, sex differences in intelligence test scores may partly be created by sampling processes and inherent biases (Dykiert, Gale, & Deary, 2009; Hunt & Madhyastha, 2008; Molenaar, Dolan, & Wicherts, 2009). A particular problem in current psychological research is that samples are often selected from an easily accessible population (often referred to as convenience or opportunity samples) rather than according to a stratified representative sampling plan. During such a recruitment process, differences between the groups may be introduced that do not naturally occur in the general population, thereby introducing artefacts. Such recruitment effects can be significant, resulting in researchers making claims about sex differences that do not actually exist in the general public. Hunt & Madhyastha (2008) suggest that it is psychometrically and ethically inappropriate to make generalisations to the wider population when studies have made use of convenience samples.

Although biases due to sampling, selection, and distributional errors may be relatively small, when considered in terms of national distributions, small differences can translate into very large differences of actual numbers of males and females, particularly at the extremes of the score distribution. An example offered by Hedges and Nowell (1995) illustrates the point very well: A small effect size (according to the guidelines offered by Cohen) of 0.3 coupled with a difference in variance of 15% would lead to 2.5 times as many men as women in the top 5% of the distribution of scores, and more than 6 times as many men in the top 0.1%.

Another psychometric issue pertains to the recency of publication of the measure employed in the research. It appears to be fairly common practice to use editions of measures that are considerably out of date which could be the reasons that significant sex differences are concluded (Neisser et al., 2006). For example, Rushton & Skuy (2006) assessed their opportunity sample of 17- to 23-year old South African university students according to the normative information for the 1993 American edition of the SPM, published 13 years previously. In light of the Flynn effect and the idea that populations are becoming increasingly intelligent by approximately three IQ points with each passing decade, it is psychometrically inappropriate to base conclusions on irrelevant normative data. It is therefore particularly important to ensure that research is making use of the most up-to-date measures and normative information possible (Brouwers, Van De Vijver, & Van Hemert, 2009).

This empirical example raises a further psychometric issue of cross-cultural test use. While the Raven's Matrices are understood to be largely "culture-free", there remain some performance differences across cultures (Brouwers et al., 2009; Raven, 2000). It therefore is apparent that in order to conduct psychometrically sound research, culturally relevant and up-to-date measures are required (Neisser et al, 1996).

Another potential cause of erroneous conclusions about sex differences is a method effect. A method effect exists when some of the differential covariance amongst a set of items is attributable to the measurement approach rather than the underlying latent ability being measured (Brown, 2006). Method effects (also referred to sometimes as 'methods factors') can arise from the modality of measurement presentation or the way a test item is presented (such as items in the form of questionnaire or multiple choice format). They may be due to the way items are worded or presented or even to the pressures of social desirability (Brown, 2006; Podsakoff, MacKenzie, Jeong-Yeon, & Podsakoff, 2003). The resulting method effect is a measurement artefact of different response styles associated with the way in which the item is presented, and not based upon different dimensions of the underlying latent factor of intelligence. According to the information presently available, there are no studies of the Raven's Matrices that assess method effects. However, an illustrative example is available from the Self-Esteem Questionnaire (SEQ).

Marsh (1996) challenged the commonly-used two-factor solution of the SEQ through the use of a single factor solution with a method effect. Traditionally, the positively- and negatively-worded items of the questionnaire are conceptualised as two distinct factors. However, Marsh identified that the patterns with which the respondents were answering the items were not based upon substantively different dimensions, but rather the covariation of the factor loadings was related to the positive and negative nature of the item wording. In light of the current dissertation and the longstanding body of literature claiming a male advantage in general intelligence, it is appropriate to consider the role of method effects as a source of potential psychometric bias.

A further psychometric explanation relates sex differences not to the differences in ability between groups of individuals but to item bias inherent within the measure. In the literature, sex differences in intelligence have generally been investigated using classical statistical methods such as t-tests or ANOVAs. However, classical test theory methods are now considered to lack the strength and specificity required to look at group differences[14] and are more appropriately addressed using modern statistical modelling methods (Embretson & Reise, 2000). Item Response Theory (IRT) and Structural Equation Modelling (SEM) provide more robust alternatives to classical methods. These methods allow for the equivalence of measurement properties across groups also known as the assessment of measurement invariance (Embretson & Reise, 2000; Keith et al., 2008). If the same score on a test of intelligence is not representative of the same level of ability in different groups as a result of different measurement properties for different groups of test takers, a test is considered to be biased (Drasgow, 1984; Horn & McArdle, 1992). This is also referred to as Differential Item Functioning (DIF). DIF refers to instances where an item on a test yields a different mean response for members of different groups with the same latent trait score (e.g. same level of intelligence). Differential Item Functioning was implicated in the findings of Abad et al., (2004), in their investigation of the APM. As discussed in section 2.3.2.1, several of the

[14] Further details about the advantages latent variable modelling techniques over classical methodologies will be detailed in Chapter 4.

Visuospatial and Verbal-analytic items showed evidence of non-uniform DIF. Of these, the Visuospatial items were found to be easier for males, while the Verbal-analytic items were easier for females. In this study, the bias was accounted for in the analyses. Yet this is clearly not the case in many other studies and could be influencing the overall results and conclusions drawn.

Wyse and Mapuranga (2009) suggest that DIF analysis plays an important role in the assurance of equity and fairness in cognitive assessments. The determination of measurement invariance at the outset of analyses ensures that any differences found between groups are in fact genuine differences rather than artefacts arising from measurement error or bias (van Der Sluis et al., 2008; Wicherts, Dolan, & Hessen, 2005). Assessment of measurement invariance is necessary for ensuring that accurate conclusions about group differences are drawn (Horn & McArdle, 1992). Without ensuring measurement invariance of a measure, it is unclear whether mean differences between groups are a genuine reflection of differences in the underlying construct, or if these differences are attributable to the bias within the measure (Horn & McArdle, 1992). A more detailed description of Structural Equation Modelling, Differential item functioning, and measurement invariance will be provided in Chapter 4.

Finally, some researchers would argue that the lack of differences in the literature can be attributed to the way in which psychometric measures of intelligence are constructed. They maintain that it is inappropriate to use standardised tests of intelligence to assess sex differences due to the nature of test construction (Halpern, In Press). When such tests are developed, the test items are balanced so as not to be biased against one sex. Unlike many would contend, the process of balancing test items is not with the intention producing equal IQ scores. Rather, it is to ensure that both sexes have equal opportunity for performing well on a test as free from biased items as possible (Mackintosh, 2001).

In light of the inconsistent findings in the literature, viewing the results from a sound psychometric perspective will allow for the findings to be better placed into context. It is clear that modern psychometric principles of test construction and analyses are the way forward in the field of intelligence research. While it is not fair to critique previous studies according to current practices, it does allow for the interpretation of findings in a more thoughtful manner.

2.4.1. Gender Differences or Similarities?

Whenever group differences are studied, it could be argued that similarities are indirectly studied as well. In a landmark article, Hyde (2005) raised this very issue in relation to sex differences in intelligence. A review of 46 different meta-analyses was conducted in an attempt to provide counter-evidence to the traditional *differences model* that argues that, psychologically, males and females are vastly different from one another. Hyde's *gender similarities* hypothesis advocates that boys and girls, men and women are similar on many, but not all, psychological characteristics and that generally they are more alike than they are different (Hyde, 2005; 2007). The hypothesis is founded on the size of effect sizes noted in the literature. Hyde claims that the vast majority of sex differences in the literature are close-to-zero ($d \leq 0.10$) or small ($0.11 < d > 0.35$) in terms of size, with very few large ($d = 0.66 - 1.00$) or very large. ($d > 1.00$).

A number of different psychological constructs typically showing sex differences in the literature were considered, including aggression, sexuality, attribution of success to ability and failure, as well spatial ability, vocabulary and the progressive matrices. On tasks of spatial abilities (Hedges & Nowell, 1995; Feingold, 1988; Linn & Petersen, 1985; Voyer et al., 1995), the review revealed effect sizes in the range $d = +0.06$ to $+0.73$[15]. On tasks of verbal abilities (Hedges & Nowell, 1995; Feingold, 1988; Hyde & Linn, 1988), the review revealed effect sizes in the range of $d = -0.40$ to -0.02. When results of the Raven's Progressive Matrices were evaluated (Lynn & Irwing, 2004), effect sizes were noted to be in the range of $d = +0.02$ to $+0.30$, which are considered to be relatively small.

Sex differences noted across the literature were found to vary considerably in magnitude according to the method of measurement (i.e., the types of tests used to assess the ability) and also according to the age of the participants. Overall, while a number of differences were identified, Hyde

[15] A positive effect size (+) represents a male advantage while a negative effect size (-) is indicative of a female advantage.

concluded that males and females are, in fact, similar on most attributes, particularly those with respect to intelligence.

In a literature that is dominated by the views that “men have more” and “men and women are different”, a review of this nature affords a refreshing perspective on the issue of sex differences. It allows some distance between the research objectives and the outcomes of the individual studies such that it is possible to consider an alternate way of interpreting the findings. There is no dispute that the gender differences model is the predominant perspective in the literature, but considering an alternative view, the gender similarities perspective allows for a more balanced and equitable evaluation of cognitive abilities in males and females. Further, one is reminded that any such differences between the sexes need not be viewed as deficiencies but as opportunities for greater understanding of what makes individuals unique (Halpern, In Press).

2.5. SUMMARY

A review of the literature of multiple cognitive abilities, although conflicted, has generally illustrated that males tend to outperform females on tasks of spatial ability, while females tend to outperform males on tasks of verbal ability. On tests of general intelligence, as measured by tests such as the SPM, the literature is even more discordant: male superiority, female superiority, and no meaningful differences are reported (Court, 1983).

It is clear from a review of this literature that the debate over sex differences in intelligence has long been unresolved and, in light of the interest in the subject, it is likely to be debated for many years to come. Extant studies in the literature of sex differences on the SPM generally have not made use of modern psychometric approaches that allow for the assessment of item bias and mean differences while accounting for measurement error. It is therefore not clear whether the male advantage often reported in the literature is due to a true male superiority in general intellectual functioning, or as a result of psychometric or statistical methodologies employed.

Modern statistical modelling methods are identified as being significantly more robust at identifying true group differences than classical statistical methodology. With advances in modern psychometric test construction methods and statistical modelling techniques, group differences in intelligence can now be assessed in a more rigorous and stringent manner. While this dissertation will, by no means, put the debate to rest, the quality of the standardisation sample and the complexity of the robust statistical analyses will make a novel contribution to better understanding the differences between males and females in general intellectual ability.

In accordance with the existing literature on sex differences in intelligence, this dissertation has four primary aims. The first aim of the dissertation is to determine whether the SPM+ is measuring the construct of general intelligence in the same way for both males and females. That is to say, is the SPM+ free from item bias, ensuring a fair assessment for both sexes? This will be determined by evaluating item characteristics and measurement properties. If the SPM+ does present equivalent measurement properties between males and females, then it is considered appropriate to proceed with the second aim of the dissertation: to determine whether there is a significant sex difference in the mean and variability of scores in the overall sample of the SPM+. Due to the large age range of participants of the standardisation sample, it is also important to consider whether age is a contributing factor to the emergence of sex differences in general intelligence. This forms the third aim of this dissertation: to assess whether sex differences emerge in younger and older groups of participants. The final aim of this dissertation is to determine whether extraneous elements inherent in the items, or method effects, are affecting performance on the SPM+.

Chapter 3 introduces the standardisation sample, procedures and measures used in the current study, while Chapter 4 explains the statistical rationale that will be employed in assessing the research aims. Specifically, Item Response Theory, in the form of Exploratory Factor Analysis (EFA) and Rasch Modelling, will be used to assess the item-level characteristics of the SPM+, followed by Multiple-group Confirmatory Factor Analysis (MG-CFA) to establish measurement equivalence and the assessment of group differences. Methods factors will be added to the analyses to determine whether attributes inherent in the SPM+ are influencing the way the participants respond to the test. Chapter 5 serves as an introduction to the finer details of the SPM+ at the item-level through a presentation of item characteristic curves and Rasch models.

This dissertation is conceptually organised into four different aims in order to effectively address the research objectives outlined. The first two objectives of this dissertation will be addressed in Chapter 6: Is the SPM+ measuring general intelligence in the same way for males and females, or is the test biased towards one group?, and Are there sex differences in the overall sample on the SPM+? The third objective of the dissertation will be addressed in Chapter 7: Are there sex differences in younger or older participants of the SPM+? Finally, Chapter 8 provides a summary of the results and limitations of the present study, offers an integrative discussion of the theoretical and practical implications of the findings, and considers possible areas for future research.

— 3 —

METHODS

3.1. INTRODUCTION

This chapter describes the methodology employed in this dissertation to investigate sex differences in the mean and variability of scores of the U.K. standardisation sample of the Raven's Standard Progressive Matrices Plus (SPM+; Raven, Court, & Raven, 2008). Details of the Raven's SPM+ and the U.K. standardisation project are first provided (section 3.2). This is followed by the sample recruitment and selection procedures (sections 3.3.1 and 3.3.1.1), and the data collection procedures (3.4). Finally, details of the final standardisation sample are provided in section 3.5.

3.2. MATERIALS

3.2.1. Raven's Standard Progressive Matrices Plus

With a long and distinguished history in the field of research, the *Ravens Standard Progressive Matrices-Plus* (*SPM+*) is a non-verbal measure of general cognitive ability suitable for children between seven to 18 years of age. The Raven's Matrices measures educative ability – the ability to perceive new relationships and patterns out of complex and novel information, and make sense and meaning (Raven, Court, & Raven, 2008). Carpenter, Just, and Shell (1990) describe the Raven's Progressive Matrices as "a classic test of analytic intelligence.... the ability to reason and solve problems involving new information, without relying on an explicit base of declarative knowledge derived from either schooling or previous experience" (p.404).

Along with its compliment measures, the Coloured Progressive Matrices (CPM; for children ages five to 11 years or those individuals with low intellectual ability; Raven, Raven, & Court, 2008); and the Advanced Progressive Matrices (APM; for individuals of high intellectual ability; Raven, Raven, & Court, 1998), the Raven's Matrices have been applied in a variety of settings. In clinical applications, the reliability of the measure and the lack of bias make it a suitable measure for neuropsychological assessment, elderly populations, or those clinical groups whose needs make the demands of more traditional cognitive assessment unrealistic. In educational settings, the Raven's has been widely used because the measure is found to be relatively unaffected by linguistic and ethnic effects. In research settings, the Matrices have often been used because of their strong theoretical background, proven research record, and its successful application across multiple languages, ethnic, and age groups.

Within the context of the literature relating to sex differences in general intelligence, the Raven's Matrices are thought to have the "highest *g* loading" of any measure of cognitive ability (Jensen, 1998, p.541) and are thought to be among the best measures of general intelligence (Jensen, 1998; Mackintosh, 1996). The Raven's Standard Progressive Matrices Plus is therefore considered the most suitable measure for answering the question of sex differences of mean and variability in general intelligence.

The SPM+ comprises 60 diagrammatic puzzles (1 practice item and 59 test items) arranged into 5 sets of 12 items per set. The items increase in difficulty in a stepwise progressive manner; items increase in difficulty within the item set, while increasing in difficulty across successive item sets. Test administrations of the SPM+ can be timed or untimed; however, during the data collection for the standardisation, the administrations were untimed. Due to the non-verbal, self-directed nature of the method of administration, the measure is well suited for both individual and group administration formats. During the standardisation data collection, both individual and group administration formats were used. For group administrations, there were no more than 15 individuals per group in each testing session. Administration procedures were the same for both the individual and groups testing scenarios.

3.2.2. Standardisation of the Raven's Progressive Matrices

The data used in this dissertation was collected as part of a larger project conducted by The Psychometrics Centre to produce a revised standardisation of the Raven's Matrices[16] for Pearson Assessment Incorporated, U.K. Prior U.K. standardisations of the Coloured, Standard, and Advanced Matrices date back to 1998.

It is considered psychometrically important to ensure that the normative data of standardised measures of cognitive ability are periodically revised in order to remain valid and reliable. Test re-norming ensures that the normative data accurately reflects the current demographic composition of the population who will ultimately be tested using the measure. In doing so, it ensures that the standard scores are a more accurate representation of ability levels of that population and the individual being tested, thereby accounting for the Flynn effect (Flynn, 2007).

It is important to specify that this study should not be considered a secondary data analysis as I was actively involved in numerous phases of the standardisation project as part of the research team at The Psychometrics Centre. From January through October 2007, I administered the SPM+ to a large number of children in Cambridgeshire. Additionally, I was involved with the recruitment of schools and examiners for participation in the project, as well as preparation of sampling matrices for the equating studies (which involved comparisons between individual to group administrations, test-retest administrations, and different versions of the tests). Further, I was involved with the selection of sample participants and assisted with the management of the sampling matrix.

Permission to use the complete standardisation data set of the Raven's SPM+ for this dissertation was kindly provided by Pearson Assessment Incorporated, U.K. All participant data remains their property and is being used in this dissertation with their written permission.

[16] The standardisation included the re-norming of the Coloured Progressive Matrices (CPM) and the Standard Progressive Matrices Plus (SPM+) for individuals 7 years and older. Re-standardisation of the Advanced Progressive Matrices (APM) was not conducted.

3.3. PARTICIPANTS

3.3.1. Recruitment

A representative sample of 489 girls and 437 boys (N = 926) between seven years 0 months and 18 years 11 months were recruited from schools and colleges across the United Kingdom. All participant sampling was conducted in accordance with a selection matrix that was representative of the 2001 U.K. population Census. This will be described in more detail in the following section.

To begin the sample selection process, primary schools, junior schools, secondary schools and colleges across the U.K. were identified according to their geographic location, the identified social standing of the catchment area, population density (i.e., urban, suburban, and rural), and ethnicities of the attending pupils. State and private schools as well as denominational and non-denominational schools were included in participant recruitment.

A total of 85 schools and colleges were selected and invited to participate in the standardisation project of the Raven's SPM+[17]. Schools were provided with information letters detailing the objectives and procedures of the standardisation (Appendix 1) and consent forms (Appendix 2). They were asked to distribute the letters and consent forms to parents and children over the age of 16 years according to predetermined sampling targets they were provided. These were determined according to the ages and year levels of the children required for the study from their school according to the census data for the school's geographical region.

Consent forms asked parents to provide detailed demographic information about the child: the child's date of birth, sex, ethnicity, use of the English language in the home, whether the child had any learning difficulties, or if the child required any learning aids (such as corrective lenses, or an in-class teaching or support assistant). The consent form also asked for

[17] A comprehensive list of participating schools and colleges is available in the *Manual of the Ravens Progressive Matrices and Mill Hill Vocabulary Scale* (Raven, Court, & Raven, 2008).

information about the parents and/or guardians of the child: highest attained level of education of all parents/guardians present in the home, the language(s) spoken in the home, and the current occupation or profession of the parent(s) or guardian(s).

Completed consent forms were collected by the schools and then returned to The Psychometrics Centre. The returned consent forms were reviewed by research staff at The Psychometrics Centre and demographic information was entered into the consent form database.

3.3.1.1. Participant selection

The sampling matrix was constructed to replicate proportions of the U.K. population according to the most recent 2001 U.K. census. This matrix was stratified in terms of five key demographic variables, each with multiple levels: 12 levels of age, two levels of gender, five levels of parent education level, four levels of race/ethnicity, and 12 levels of geographic region. Each of these variables will be described further in section 3.5. The sampling matrix was used to guide both the initial sample recruitment and the final selection of participants from the SPM+ consent form database according to the five key demographic criteria indicated above. The final sample included only those children who were fluent in English, regardless of any additional languages spoken in the home, and those who did not require additional special needs support.

3.3.2. Data Protection and Ethical Considerations

In accordance with data protection protocol, the consent form database was password protected and accessible only to those members of The Psychometrics Centre working on participant selection for the sampling matrix. In order to further protect the privacy of participants, original paper copies of all consent forms were locked in filing cabinets in a secure room.

Prior to the commencement of participant recruitment and data collection for the Raven's Matrices standardisation project ethical approval was obtained by The Psychometrics Centre from the University of Cambridge Psychology Ethics Committee. All research undertaken complied fully with ethical guidelines outlined by the British Psychological Society.

3.4. Data Collection

Data collection of the sample took place between November 2006 and January 2008. Administration of the SPM+ was conducted primarily by educational psychologists who had prior experience with administration of educational assessments. Additional specialised training was provided to all examiners relating to the administration of the SPM+ for the standardisation to ensure that all of the protocols and procedures were followed correctly. Upon successful completion of the SPM+ training, examiners were provided with lists of participants to be tested. Participants were identified by name, school name, SPM+ identification number, and the format of administration the child was to be tested (i.e. individual or group administration). In order to preserve the participant's privacy, further communication about the participant (i.e., requests for re-testing, identification on record forms) used identification numbers rather than names.

Both individual and group administration methods were used in the standardisation data collection to ensure that both methods of administration produced equally valid and reliable results. It was the responsibility of the examiners to contact the schools of the identified children and to arrange appointments for test administration directly with the schools. Upon completion of the test administration, the completed test protocols were returned to The Psychometrics Centre for data processing, scoring and inclusion in the protocol database.

3.4.1. Standardisation Test Administration

The procedure for administration was the same for both individual and group administrations of the SPM+. Schools were asked to provide rooms that were suitable for test administration: quiet, well-lit and that had individual workspaces for the participants, such as an unused classroom or a preparation room. Rooms such as a library were deemed unsuitable due to possible distractions to the test takers.

Prior to the participant's arrival to the testing room, testing materials were placed on each desk: a record form, a test item booklet, and a pencil without a rubber. To ensure protection of personal information, the child's record form was marked with their predetermined identification number rather than their name. Upon arrival to the testing room, children were asked to sit at their workspace, not to look at the testing materials and wait to be instructed how to proceed.

When the participant was ready (or all participants were ready in the case of a group administration), the examiner read aloud the administration instructions verbatim from a script provided in the standardisation testing manual (see Appendix 3 for full administration instructions). It is psychometrically important that all instructions provided to participants for a standardisation are equivalent across participants and testing sessions. This ensures greater testing reliability and accuracy in the testing procedures.

In addition to the verbal instructions provided to test participants, the examiner illustrated how to complete a test item using a poster-sized example at the front of the room. The children were then asked to complete a practice item. They were also given the opportunity to ask questions in order to clarify any concerns they had with the task or the testing procedure. Upon resolution of any queries, children were instructed to open their test booklet and begin completing items as quickly and carefully as possible. The administration for the standardisation was not timed, and was therefore completed once the child had completed all of the items on the assessment or once they had reached a series of items they could no longer answer and had stopped attempting further items.

Once the testing session was complete, record forms were sent to The Psychometrics Centre whereupon they were assessed for completion, and entered into the protocol database. All items were coded into the data base as correct (1), incorrect (0), or not answered (99). As is convention within the literature, item scores were converted to a binary metric, where "1" represents a correct item score, and "0" represents either an incorrect item or an item not answered (R. Lynn, personal communication, 2009). Analysis of missing data is not within the scope of the current dissertation, and will be discussed further in Chapter 4.

To ensure data protection protocols were upheld, record forms were entered into the database according to the child's unique identification number rather than their name. This identification number was then cross-referenced with the master list of examinees and the standardisation sampling matrix. Once the record forms were entered into the protocol database, they were moved to locked filing cabinets and stored in a secure room at The Psychometrics Centre.

3.5. Final Standardisation Sample

The final standardisation sample comprised 926 children between the ages of seven years 0 months and 18 years 11 months (7:0-18:11; Table 1) and reflects proportions of the population according to the 2001 U.K. Census upon which the standardisation sampling matrix was constructed. The final sample will now be described in terms five key demographic variables: age, sex, parent education level, ethnicity, and geographic location.

3.5.1. Age and Sex

The final standardisation sample comprised 926 children (437 males, 489 females). Children were grouped into 12 age categories (year:month): 7:0-7:11, 8:0-8:11, 9:0-9:11, 10:0-10:11, 11:0-11:11; 12:0-12:11, 13:0-13:11, 14:0-14:11, 15:0-15:11, 16:0-16:11, 17:0-17:11, and 18:0-18:11 (Table 1; Mean age=12.33 years, S.D.= 3.284; Mean age_{male} = 12.00, S.D. = 3.222; Mean $\text{age}_{\text{female}}$ = 12.44, S.D. = 3.329). Frequencies of male and female participants are provided by year in Table 1 and by groups of younger and older participants in Table 2.

Table 1. Details of Age and Sex of the SPM+ Standardisation Sample

Age (Years : Months)	Frequency (N)	Percent of Total Sample (%)
7:0 – 7:11	60	6.5
Male	29	6.6
Female	31	6.3
8:0 – 8:11	95	10.3
Male	49	11.2
Female	46	9.4
9:0 – 9:11	87	9.4
Male	44	10.1
Female	43	8.8
10:0 – 10:11	88	9.5
Male	40	9.2
Female	48	9.8
11:0 – 11:11	74	8.0
Male	40	9.2
Female	34	7.0
12:0 – 12:11	90	9.7
Male	51	11.7
Female	39	8.0
13:0 – 13:11	98	10.6
Male	41	9.4
Female	57	11.7
14:0 – 14:11	71	7.7
Male	29	6.6
Female	42	8.6
15:0 – 15:11	57	6.2
Male	32	7.3
Female	25	5.1
16:0 – 16:11	80	8.6
Male	34	7.8
Female	46	9.4
17:0 – 17:11	83	9.0
Male	27	6.2
Female	56	11.5
18:0 – 18:11	43	4.6
Male	21	4.8
Female	22	4.5
Total	926	100.0
Male	437	47.2
Female	489	52.8

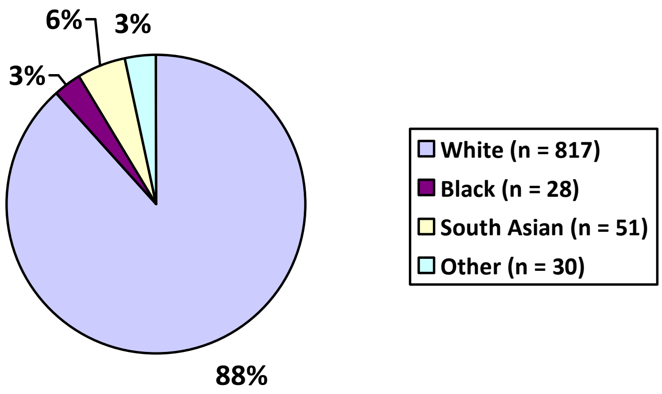
Table 2. Details of Age Groups by Sex of the SPM+ Standardisation Sample

Age (Years : Months)	Frequency (N)	Percent of Age Group Sample (%)	Percent of Total Sample (%)
7:0 – 14:11	663	100	72.0
Younger Male	323	48.7	38.9
Younger Female	340	51.3	36.7
15:0 – 18:11	263	100	28.4
Older Male	114	43.3	12.3
Older Female	149	56.7	16.1

3.5.2. Ethnicity

The ethnicity of each child in the sample was identified by his/her parent or guardian (or by themselves if aged 16 or older). Ethnic groups included in the standardisation were: *White*, *Black* (including *African*, *Caribbean*, and *Other*), *South Asian* (including *Indian*, *Pakistani*, *Bangladeshi*, and *Other*), and *Other* (including *Chinese* and mixed race).

Figure 4. Details of Ethnic Groups in the SPM+ Standardisation Sample



3.5.3. Level of Parental Educational Achievement

The standardisation sample was stratified according to five levels of parental educational achievement reflecting the number of years of education completed by the parents (Table 3). When the father was living with the child, the father's education level was used. If the father was not present in the home, the mother's (or guardian's) level of education was used.

Table 3: Details of Parental Educational Achievement of the SPM+ Standardisation Sample

Parent Education Level	Description	Frequency (N)	Percent of Total Sample (%)
1	No recognised educational qualifications	146	15.8
2	Up to five GCSE A-C or equivalent	193	20.8
3	Five or more GSCE A-C or equivalent	225	24.3
4	Two or more A Level or equivalent	95	10.3
5	University degree or equivalent	267	28.8
Total		926	100.0

3.5.4. Geographic Region

For the purposes of the standardisation, the United Kingdom was divided into 12 geographic regions as identified by the 2001 Census report: East Midlands, East of England, London, North East, North West, Northern Ireland, Scotland, South East, South West, Wales, West Midlands, and Yorkshire & Humberside (Table 4).

Table 4: Details of Geographic Region of the SPM+ Standardisation Sample

Geographic Region	Frequency (N)	Percent of Total Sample (%)
East of England	50	5.4
East Midlands	90	9.7
London	74	8.0
North East	36	3.9
Northern Ireland	61	6.6
North West	117	12.6
Scotland	120	13.0
South East	92	9.9
South West	89	9.6
Wales	31	3.3
West Midlands	67	7.2
Yorkshire & Humberside	99	10.7
Total	926	100.0

3.6. Summary

Using the U.K. standardisation sample of the Raven's SPM+ as a way to investigate sex differences in mean and score variability in general intelligence affords many opportunities, namely the opportunity to use a large sample that is representative of the current population in the United Kingdom. The Matrices have a long and distinguished history in the assessment of cognitive ability (Jensen, 1998; Mackintosh, 1996) ensuring that reliable and valid conclusions may be drawn. Within the existing literature pertaining to sex differences in general intelligence as measured by the Ravens Matrices, few studies have employed representative samples. Further, few studies have availed themselves of advances in psychometrics that allow identification of bias in test items and the reliable evaluation of group differences. The current dissertation aims to address these identified weaknesses by using a representative sample of the U.K. population and to employ advanced psychometric techniques to appropriately address the issue of sex differences in mean and variability of general intelligence. Data will be analysed using structural equation modelling in Mplus version 5.21 (Muthén & Muthén, 2009). These statistical methods are considered to be robust and reliable techniques for exploring potential item bias, assessing latent factors, and examining group differences while accounting for measurement error in the analyses (Brown, 2006; Camilli & Shepard, 1994). The advanced statistical modelling theory and methods will now be detailed in Chapter 4.

— 4 —

Statistical rationale of Analytic Techniques

4.1. Introduction

When faced with the question of sex differences in intelligence, the results are largely dependent upon the way that intelligence is conceptualised, measured and evaluated (Halpern & LaMay, 2000). In the literature, sex differences in intelligence have generally been investigated using classical methodology such as t-tests or ANOVAs. However, classical test theory methods are now thought to lack the strength and specificity required to investigate group differences (Embretson & Reise, 2000), and are more appropriately assessed using modern statistical modelling methods. Item Response Theory (IRT) and Structural Equation Modelling (SEM) provide more robust alternatives allowing for the investigation of relationships between observed (i.e., measured) variables and latent (i.e., unmeasured) factors, while accounting for measurement properties and error in the analyses (Embretson & Reise, 2000).

This chapter provides theoretical descriptions of the statistical modelling techniques that will be practically applied in the analysis of the Raven's SPM+ data that will be reported in results Chapters 6, 7, and 8. First, the approach to analysing the SPM+ data at the item level will be described (section 4.2). Item Response Theory (IRT) will then be described in relation to its counterpart, Classical Test Theory. Next, Latent Variable Analysis will be discussed including a review of both Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) techniques (section

4.3). A discussion of group-level analyses will be then be provided (section 4.4). A discussion will be provided of the importance of establishing equivalence in measurement properties between groups (also known as *measurement invariance* or *Differential Item Functioning*) prior to the evaluation of group differences. Multiple-Group Confirmatory Factor Analysis (MG-CFA) will then be described as a method of assessing group differences (section 4.5). Model specification and assessment of model fit will then be discussed. Finally, a summary will be provided offering further evidence that SEM is the optimal method for assessing group differences in latent factors of intelligence while acknowledging critiques of these modelling techniques (section 4.6).

4.2. Item Response Theory

Item Response Theory (IRT) is now recognised as one of the most appropriate methods for assessing item-level test data (Embretson & Reise, 2000). While both Classical Test Theory (CTT) and IRT conceptualise an individual’s response to a test item as related to the underlying latent construct, they model the association between the item and the latent construct in different ways. Latent constructs are those which are not directly measurable or observable but are inferred through the measurement and observation of other variables. Within the context of the current discussion of sex differences in intelligence, a latent trait can be thought of as the individual’s level of cognitive ability being measured by the items on the Raven’s SPM+. The psychometric properties of a scale of items is summarised in CTT by a single omnibus statistic, such as Cronbach’s α which is based upon correlations between different items on a test. In contrast, IRT represents the variation in endorsing an item as a function of the respondent’s level of the latent construct in relation to the item parameters or characteristics (Baker, 2001; Santor, Ramsay, & Zuroff, 1994).

One central concept within IRT is the item characteristic curve (ICC). An ICC plots a non-linear regression of the probability of responding correctly to an item as a function of the latent construct, often denoted by θ in psychometric literature (Crocker & Algina, 2006; Hulin, Drasgow, &

Parsons, 1983). In most applications of IRT, the item characteristic curve is assumed to have an ‘S’ shape, as illustrated in Figure 5, and is often referred to as the *normal ogive*.

Figure 5: Normal Ogive Curve or the 1-Parameter Logistic Model

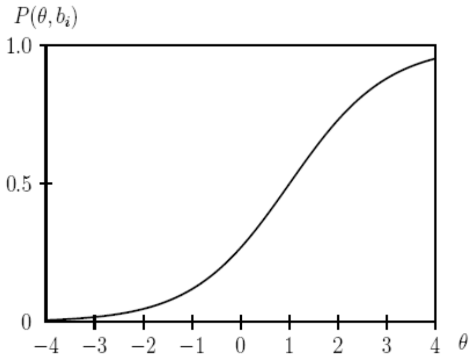


Figure 5 illustrates that as the score of the latent trait (or ability represented by θ) increases across the horizontal, so does the likelihood of a correct response, as indicated by the probability along the vertical axis. The measurement metric of the latent trait scale can be anywhere between -∞ and +∞, but to make it more interpretable the convention is to select the unit of measurement such that the mean latent trait score is 0 and the standard deviation is 1 for the population of interest (Crocker & Algina, 2006). With respect to ability, the metric most often used is -4 to +4 (Partchev, 2004).

There are several important properties of the normal ogive ICC:

1. The curve rises continually from left to right, and is said to increase monotonically;
2. The lower and upper asymptotes approach the limits, but never quite reach 0 or 1 respectively; and
3. The normal ogive curve is directly related to the normal distribution curve, where the proportions of correct responses are expressed as functions of z-scores (Crocker & Algina, 2006).

The ICC can be represented symbolically by the following equation:

Equation 2:

$$P_i(\theta) = c_i + (1 - c_i)\frac{1}{\left[1 + \exp\left\{-Da_i(\theta - b_i)\right\}\right]}$$

where:

- θ represents the value of the latent trait (e.g., cognitive ability),
- P(θ) represents the probability of a positive response,
- D is a scaling constant equal to 1.702, and
- a, b, and c are the parameters characteristics of an item.

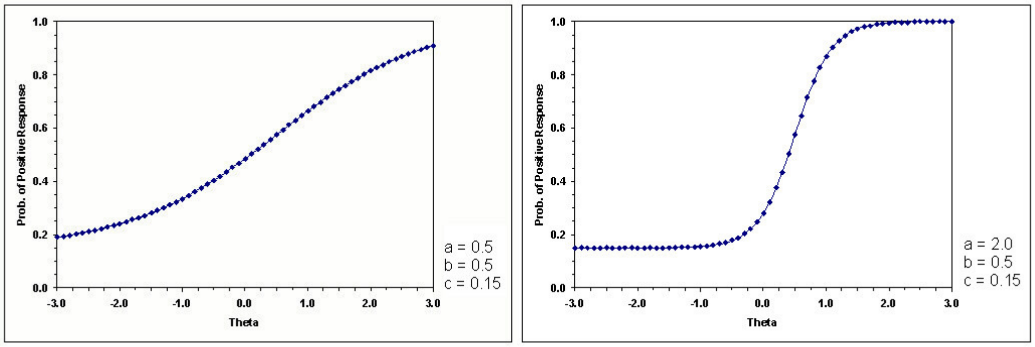
When interpreting these curves, one can think of having subgroups of participants at each point along the horizontal ability scale, each of whom have the same latent trait score (i.e., the same level of cognitive ability), with each member of the sub-group having equal likelihood of getting the item correct. With this in mind, an ICC can be described in terms of three parameters: (a) discrimination, (b) difficulty, and (c) a guessing parameter, details of which will now be provided.

4.2.1. Discrimination Parameter

The discrimination of an item describes how well an item can distinguish between participants of different levels of the latent trait of cognitive ability. Discrimination directly affects the slope or steepness of the item characteristic curve in its middle section (Figure 6). The steeper the item curve, the better the item can discriminate between members of low and high ability groups. The flatter the curve, the less the item is able to discriminate between levels of the latent trait since the likelihood of correct response at the low ability levels is similar to the likelihood at the higher ability levels (Baker, 2001).

On the left of Figure 6, a curve with a relatively flat slope (e.g. a = 0.5) represents an item that is not very effective at differentiating between different levels of the latent trait or between individuals with different levels of ability.

Figure 6: Low versus high item discrimination

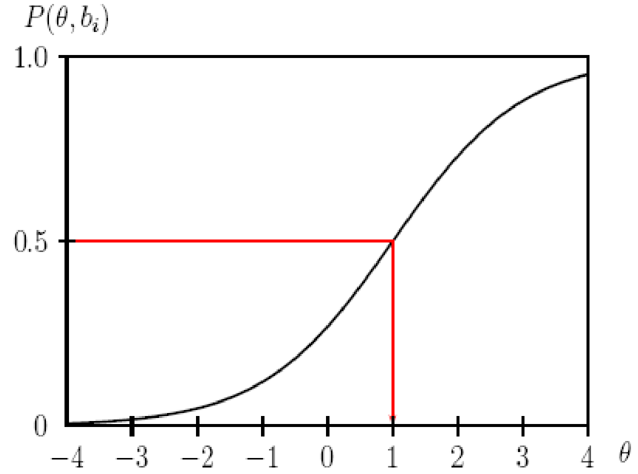


In contrast, on the right of Figure 6, a curve with a steep slope (e.g. a = 2.0) represents items that are able to discriminate clearly between individuals of different levels of the underlying latent trait or ability. An item with optimal discrimination will be graphically represented by a curve with a moderately increasing slope, as seen in Figure 5.

4.2.2. *Difficulty Parameter*

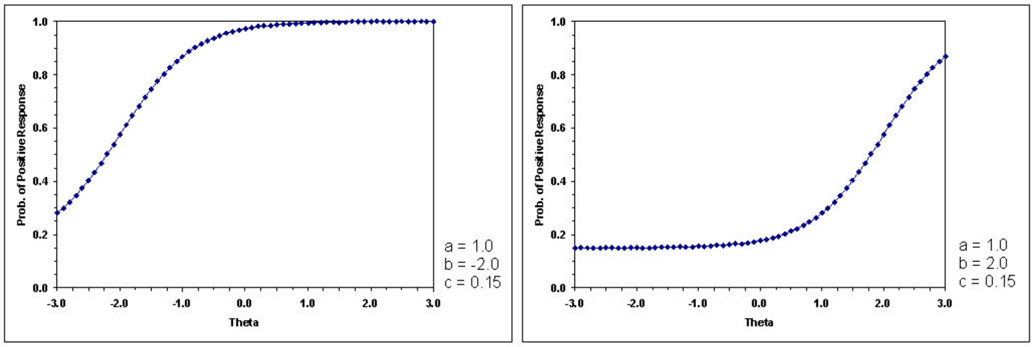
The difficulty of the item (also known as the 'item location') corresponds to the value of the latent trait at the point where the predicated probability of a respondent correctly endorsing an item is 50% (Figure 7).

Figure 7: Locating the difficulty of an item on the ability / difficulty axis (Partchev, 2004)**.**



As can be seen in Figure 8, the location of the item response function is directly affected by the item difficulty parameter.

Figure 8: Low versus high item difficulty



Items with low difficulty values, as illustrated on the left of Figure 8, are those items that are frequently answered correctly, even by those individuals who are low on the latent trait. Conversely, items that are not frequently answered correctly, even by individuals who are high on the latent trait, possess high difficulty values as illustrated on the right of Figure 8 (Crocker & Algina, 2006; Sharp, Goodyer, & Croudace, 2007).

In relation to the item discrimination, when the discrimination of an item is high (i.e., the item curve shows a steep slope), the item provides more information about the latent trait, with the item information concentrated around the item difficulty. Items with low discrimination ability (i.e., the item curve shows a shallow slope) are less informative with the item information scattered along a greater range of the latent trait. The item difficulty can also be conceptualised as equivalent to the item-to-total correlation or the proportion-correct score (“p-value”) in Classical Test Theory. These are inversely proportional with larger values of *p* or smaller values of difficulty (*b* in Equation 2) indicating easy items (Jones, 2009).

4.2.3. Pseudo-guessing Parameter

The third parameter for describing an item characteristic curve is the ‘pseudo-guessing’ parameter or the ‘lower asymptote’ parameter. It indicates the likelihood of a correct response for individuals with a very low value of the latent construct (θ) in relation to a positive value of 1/β, where β is the number of response categories in an item with multiple-response options. It has been suggested that individuals of very low ability will likely use an answering strategy that involves random guessing. Random guessing enables them to select the correct response with a probability of 1/β. This parameter is not relevant for the current dissertation because, as is convention in the Raven's literature, it will be assumed that the effects of guessing will be equivalent across individuals, and hence will be held to equality in the analyses. Further information about the pseudo-guessing parameter will not be provided here, but is available in Crocker and Algina (2006).

While the normal ogive curve is useful for general theoretical understanding, in terms of practical applications, IRT employ different logistic models that account for the different item parameters. Two of these will now be described: the One-Parameter Logistic Model (also referred to as the 1PL or the *Rasch* Model) in which item difficulty is used to predict the likelihood of a correct response, and the Two-Parameter Logistic Model (also known as the 2PL) in which both item difficulty and discrimination are used in the prediction of the probability of a correct item response.

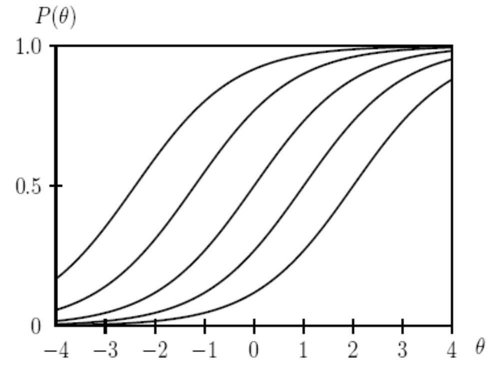
4.2.4. The One-Parameter Logistic (1PL) Model

The simplest IRT model for dichotomous items (as is the case with the SPM+ items which are scored as either correct or incorrect) is known as the one-parameter logistic model (1PL). It is also known as the Rasch model (Rasch, 1966). In reference to Equation 2, the 1PL model has only one freely estimated parameter, difficulty. The discrimination parameter is held to a constant across items (i.e., a = 1) and the pseudo-guessing parameter is held to zero (i.e., c = 0).

In this model, the proportion of individuals responding correctly is directly related to their level of the latent trait in relation to the difficulty of the item. Another way of thinking of the Rasch model is that all items on the test are said to have equal power to discriminate between individuals of low and high levels of the latent trait. As such, if all items of a test are plotted simultaneously, each item curve will display the same slope but will be located at different points along the continuum of ability in relation to their determined difficulty level (Figure 9).

Figure 9: Item Characteristic Curves of five items in a 1PL model

(Partchev, 2004).



In this example, five items of varying difficulty are shown to have equal slopes, lying parallel to one another but never crossing. With respect to the evaluation of items on a test of intelligence, such curves could be important in confirming that test items increase successively in terms of difficulty across the test which is generally the aim of such tests. Details of the Rasch analyses of the Raven's SPM+ standardisation data are provided in chapter 5.

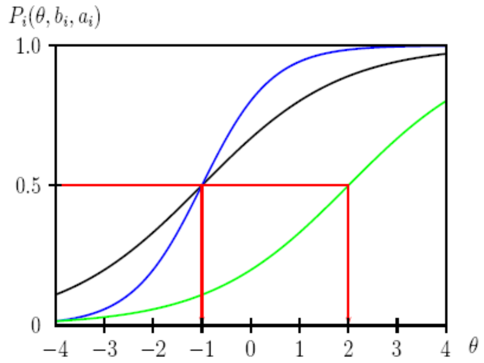
4.2.5. The Two-Parameter Logistic (2PL) Model

The second form of logistic model discussed in this dissertation is the two-parameter logistic model (2PL), also known as the Confirmatory Factor Analysis model for binary data. The 2PL is an extension of the Rasch model that allows both the discrimination and difficulty parameters of each item to be freely estimated, while the guessing parameter is held to a constant. In reference to Equation 2, a 2PL model is obtained by setting c = 0 while allowing a and b to be freely estimated.

Examining the ICCs of the 2PL model allows us to see a visual representation of the relationship between item parameters and the levels of the latent trait. Unlike the curves seen in a Rasch model (Figure 9), the curves in a 2PL model will differ with respect to both the slope of the curve and their location along the continuum of latent trait. Figure 10 presents three different item characteristic curves to illustrate this concept.

Figure 10: Item Characteristic Curves of 3 items in a 2PL model

(Partchev, 2004)



The blue and black curves have the same difficulty level, or the point at which the ability level yields a 50% chance of answering the item correct. However, the blue curve has a steeper slope than the black, indicating a greater discrimination parameter. Similarly, the green curve illustrates the same slope as the black curve, but as it is representing a more difficult item, the green curve is located farther to the right on the x-axis. It is also important to note that ICCs in a 2PL may cross one another when plotted together, as is the case with the blue and black curves to the left of the figure. This illustrates that, for examinees of different levels of the latent trait, the item may differ in terms of difficulty. The item represented by the black curve is more difficult for individuals of high ability, while the item represented by the blue curve is more difficult for individuals of lower ability (Partchev, 2004). Details of analyses with the Raven's SPM+ data using a 2PL methodology will be used to assess a 1-factor model of latent mean score differences for multiple groups by sex in Chapter 6, and to assess a 1-factor model for multiple groups by sex and age in Chapter 7. Strategies for assessing the Raven's SPM+ data at the group level will now be discussed.

4.3. Latent Variables & the Common Factor Model

First described by Spearman (1927), factor analysis has become one of the most commonly used multivariate statistical procedures in applied research. If a collection of observed variables (also known as indicators in the modelling literature) are thought to be related in some way, they are said to share an underlying, unobservable latent factor that may account for this inter-relationship. When considering a collection of related indicators (such as a sub-set of items on an intelligence test), factor analysis techniques can be employed to attempt to account for the variation and covariation among these observed indicators by some common factor. If this underlying latent factor were partialed out, the inter-correlations among the observed indicators would be zero (Brown, 2006). Consequently, factor analysis serves to understand the covariation among a collection of observed indicators in the simplest, most parsimonious manner.

According to the common factor model (Brown, 2006; Thurstone & Thurstone, 1941), each indicator in a set of observed variables is a linear function of one or more common factors and one distinctive factor. Consequently, in factor analysis the variance (or the amount of variability in participant responses) of each indicator is partitioned into two components: (1) *common variance*, which is the variance accounted for by the latent factor which is an estimate of the variance shared with other indicators in the analysis; and (2) *unique variance*, which is the reliable variance specific to an indicator along with additional measurement error known as random error variance. The error variance reflects the variance unaccountable by the latent factor. In relation to preceding sections of this chapter, factor analysis with binary outcomes can be thought of as equivalent to a two-parameter normal ogive IRT model (Brown, 2006; Field, 2005).

There are two methods of factor analysis serving different purposes: Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA). While each of these methods aims to reproduce relationships observed among a group of indicators with a more parsimonious set of latent variables, EFA and CFA differ by the nature and number of *a priori* specifications placed upon the model (Brown, 2006). Each of these techniques will now be described.

4.3.1. Exploratory Factor Analysis

Exploratory Factor Analysis (EFA) is a data-driven approach which serves as an exploratory or descriptive method for determining the appropriate number of common latent factors for a set of indicators (Brown, 2006). As such, the researcher does not impose any specifications as to the number of underlying factors or about the relationships between common factors or variables that exist within the data. Consequently, EFA is often used as an exploratory or descriptive technique in the early phases of analysis towards the specification of the most appropriate model in Confirmatory Factor Analysis (CFA).

It is important to note that EFA is fundamentally distinct from its counterpart, Principal Components Analysis (PCA), which often is mistakenly used in the place of EFA. The primary goal of PCA is data reduction – to reduce a large number of indicators to a smaller set of variables that account for the large amount of observed variance (Kashy, Donnellan, Ackerman, & Russell, 2009). PCA is not considered robust or sensitive enough to fully account for the underlying construct in the current measurement model because it does not differentiate between common and unique variance. Therefore, EFA techniques were chosen over PCA for this dissertation.

Primarily, there are three phases involved in Exploratory Factor Analysis as they relate to this thesis: 1) factor selection; 2) factor extraction; and 3) factor rotation. Each of these will now be described.

4.3.1.1. Factor Extraction

A number of different methods can be used to estimate the common factor model, of which the most commonly used is maximum likelihood (ML). However, due to the categorical nature of the data of the SPM+, a robust weighted least means squares estimator (WLSMV) is used in Chapter 5 for factor extraction. The WLSMV is a full information estimator that assesses how well the factor solution is able to replicate the relationships among the indicators in the input matrix (Brown, 2006). Another key advantage of this estimator is that it also provides a complement of fit indices to guide further model specification.

It is important to note that the number of factors that can be extracted by EFA is limited to the number of observed indicators that are in the analysis. Using a WLSMV estimator the number of parameters estimated by the factor solution (*a*) must be equal to or less than the number of indicators (*b*) in the input correlation matrix (i.e., $a \leq b$). For the current sample, the case per item ratio is 16, which is sufficient for the stability of factors (Costello & Osborne, 2005).

4.3.1.2. Factor Selection

After the appropriate estimator has been determined, the most suitable number of factors must be determined that can be used to explain the data. EFA techniques are based upon the concept of eigenvalues which can be thought of as representing the variance of the indicators explained by the identified factors of the model (Brown, 2006). In assessing the most appropriate factor solution from the EFA results of the SPM+ data, two procedures based upon eigenvalues were used for factor selection: 1) the Kaiser-Guttman rule; and 2) the scree test.

In the Kaiser-Guttman rule, also known as the ‘eigenvalues > 1.0’ rule, the researcher determines how many eigenvalues greater than 1.0 are derived in the correlation matrix. The number of eigenvalues greater than 1.0 is then used to determine the number of latent constructs that exist within the data. The rationale behind this rule is that, because eigenvalues represent variance, if an eigenvalue is less than 1.0, then the corresponding factor accounts for less variance than the indicator itself (Brown, 2006; Field, 2005).

In the scree test, eigenvalues are plotted against the suggested factor structure. The point at which there is a substantial decline in the magnitude of eigenvalues is seen as the point where the most suitable number of factors has been reached (Brown, 2006).

4.3.1.3. Factor Rotation

Once an appropriate factor structure of the data has been determined, the factors are rotated to increase the interpretability by maximising factor loadings closer to 1.0 and minimising loadings close to 0.0. It is important to mention that rotation does not apply for factor solutions with fewer than 2 factors. Rotation of the factors is a mathematical transformation and does not modify the fit of the model solution in any way.

Once rotated the factor loadings of the indicators are examined to identify their primary loadings, and in some instances, any cross-loading or secondary loadings that may exist. It is popular convention to interpret factor loadings greater than or equal to 0.30 or 0.40 as *salient*, or that the indicator is meaningfully related to a primary or secondary factor (Brown, 2006; McDonald, 1999).

Two types of rotation are possible: *orthogonal* and *oblique*. In orthogonal rotations, the factors are constrained to be uncorrelated, while in oblique rotation the factors are allowed to correlate with one another. In the EFA models presented in Chapter 5, a geomin oblique rotation was used. Oblique rotations are often preferred over other methods of rotation because they provide a more authentic representation of how factors are interrelated. Further, when the EFA is used as a precursor to a CFA, an oblique rotation is more likely to generalise than an orthogonal solution (Brown, 2006).

In the current dissertation, EFA analyses were conducted in order to validate previous findings stated in the literature and to inform the factor structure of the subsequent empirical models. While the underlying factor structure of the Ravens has been extensively studied and debated, there has yet to be a consensus reached. Despite being considered one of the best measures of the unidimensional construct of general intelligence, *g*, (Jensen, 1998; Raven, Raven, & Court, 1998c), other researchers have argued for the existence of multiple factors in the Ravens matrices (Lynn, Allik, & Irwing, 2004; van der Ven & Ellis, 2000; DeShon et al., 1995). Due to the lack of consensus in the literature regarding the underlying factor structure of the Raven's, it was deemed appropriate to begin analyses in the current dissertation with an assessment of dimensionality. Chapter 5 provides details of the EFA findings from the UK standardisation of the Raven's SPM+.

4.4. Confirmatory Factor Analysis

CFA is a form of Structural Equation Modelling (SEM) that pertains to the relationships between observed variables or *indicators* (e.g. test items, questionnaire ratings) and latent variables or *factors* (Brown, 2006). Unlike its counterpart EFA, CFA is hypothesis-driven, where all aspects of the CFA model are specified by the researcher. The researcher must have a firm theory and evidence basis for the number of factors that exist within the data, and to which extent the factors are related. Due to its basis in *a priori* empirical and conceptual grounds, CFA has many advantages over EFA.

First, the resulting CFA is testing a much more parsimonious model by specifying the number of factors, the factor loading patterns, and a suitable error theory (such as random or correlated error variances; Brown, 2006). Second, the measurement error is accounted for separately from the indicator in the determination of the latent construct. Many classical methods (such as regression analyses or correlation) make the assumption that the data is free from error, which is rarely the case in social sciences. In CFA, the error variances are excluded so that the resulting relationships between variables can be estimated after measurement error has been adjusted. Further, it is possible to specify error covariances (i.e. correlations between error variances of the indicators) within a CFA. Generally the underlying assumption is that an observed relationship between two indicators loading on the same factor is due entirely to the latent dimension. This may not always be the case. There are instances where method effects introduce additional common error covariance not explained by the latent construct, which, in an EFA may produce method factors without substantive basis.

For example, a questionnaire relating to self-concept that undergoes EFA might produce a 2-factor solution accounting for positively and negatively worded items that are evaluated as being substantively meaningful (e.g. “positive self-concept” vs. “negative self-concept”). However, once subjected to investigation by CFA allowing the error variances to correlate, it could be argued that the differential covariance among the items is not based upon substantively different dimensions, but rather due to a method effect of the item wording (Brown, 2006; Marsh, 1996). The ability to specify correlated errors is identified as a particular strength of CFA (Brown, 2006).

A third advantage of CFA, particularly relevant to this dissertation, is the ability to assess measurement invariance, or the equivalence of measurement models across groups. Measurement properties must be equivalent in subgroups of the population, otherwise the test is considered to be bias. This means that the same score on the latent construct is not representative of the same level of ability in different group members. It is, therefore, important and necessary for the mean comparison between groups of individuals. The determination of measurement invariance ensures that any differences found between groups are in fact genuine differences rather than artefacts arising from measurement error or bias. Measurement invariance within the context of group differences will now be discussed further.

4.4.1. Measurement Invariance and Group Differences

Meaningful comparisons of mean differences across groups (e.g. ethnicity, gender, culture) require that measurement equivalence holds. Measurement equivalence exists when the relation between the observed test scores and latent attribute are identical across subpopulations of test-takers (Drasgow, 1984; Horn & McArdle, 1992). Violations of this indicate measurement non-equivalence, or Differential Item Functioning (DIF).

DIF refers to instances where an item on a test yields a different mean response for members of different groups with the same latent trait score (e.g. intelligence). Horn and McArdle (1992) argue that "...evidence of invariance of measurement is necessary for drawing clear inference from results" (p. 118), while Wyse and Mapuranga (2009) suggest that DIF analysis plays an important role in the assurance of equity and fairness in cognitive assessments. Without ensuring measurement invariance of a measure, it is unclear whether mean differences between groups are a genuine reflection of differences in the underlying construct, or if these differences are attributable to the bias within the measure (Crane, Belle, & Larson, 2004; Crane, Gibbons, Jolley, & van Belle, 2006; Crane et al., 2007).

To illustrate the importance of establishing measurement invariance, let us consider the following example comparing the increase of verbal comprehension by age in adulthood by comparing average intelligence scores for groups of adults of different ages. In this instance it is necessary to be certain that verbal comprehension is being measured in the same way in all age groups. If measurement invariance were not established, any mean differences could be falsely explained as differences in verbal comprehension when in actuality two different constructs could have been measured (for example, verbal comprehension in one instance and verbal fluency in another).

Another example of the importance of the establishment of measurement invariance is the 'Flynn Effect' (Flynn, 1987; 2007; 2009), a well-documented theory that IQ increases across approximately 3 IQ points per decade. Wicherts et al. (2010; 2005) argue that the increases noted by Flynn may be due to the way that intelligence has been measured across time, not necessarily due to the underlying construct of intelligence itself.

In order to assess measurement invariance across groups, constraints must be placed upon parameters of the measurement model to examine the equivalence of the measurement properties. The measurement model relates to the measurement characteristics of the indicators (observed measures), thereby consisting of the related factor loadings, the indicator intercepts, and residual error variances (Brown, 2006). With respect to the current dissertation, the assessment of measurement invariance in group comparisons of intelligence is particularly important; for example, do the items on the Ravens SPM measure the same constructs (i.e. the same factor structure) and demonstrate equivalent relationships to these constructs (i.e. equal factor loadings) for both males and females? Are there items on the SPM+ that are biased towards males or females; that is to say, do these items yield significantly higher or lower scores for one group despite equivalent levels of general intelligence (Brown, 2006)? These questions serve as one of the central concepts of this dissertation, and will be discussed further in Chapters 6 and 7.

4.4.1.1. Population Heterogeneity

Upon the determination of measurement invariance, it is also possible to assess the structural parameters of the measurement model through an evaluation of the factor variances, covariances, and latent means. An examination of the group concordance of these structural parameters is considered one method of testing population heterogeneity; that is, is there variation across groups in the dispersion, interrelationships, and levels of the latent factors? Tests of equality of means can be considered comparable to t-tests or ANOVAs, but offer advantages over classical test methodology as the comparisons are made within the context of the measurement model, having been adjusted for measurement error and correlated residuals (Brown, 2006).

Two methods are recommended for the evaluation of multiple-group comparisons CFA measurement models: 1) Multiple Indicator Multiple

Causes (MIMIC) modelling; and 2) Multiple-Groups CFA (MG-CFA). Although MIMIC is proven to be the more parsimonious method for assessing multiple-group measurement models, this method is limited in its ability to only assess two potential sources of measurement invariance: factor means and indicator intercepts. The other measurement and structural parameters of the model (i.e. factor loadings, error variances and covariances, factor variances and covariances) are assumed to be equivalent across groups at all levels. In light of the aims of this dissertation, such assumptions are inappropriate, and therefore, MG-CFA methods will be used in this dissertation for the evaluation of measurement invariance and group differences in the Raven's SPM+ as they allow for all aspects of measurement invariance and population heterogeneity to be assessed.

4.5. Multiple-Groups Confirmatory Factor Analysis (MG-CFA)

In order to assess all aspects of measurement invariance and population heterogeneity (i.e., factor loadings, intercepts, residual variances, factor variances, factor covariances, and latent means) for multiple groups, multiple-groups confirmatory factor analysis (MG-CFA) methods can be used. Being able to assess the full compendium of measurement properties is considered a particular advantage of MG-CFA methods over the MIMIC model (Brown, 2006).

It is possible to conduct simultaneous CFA analyses for more than one group (e.g., males and females) using MG-CFA. For each of the groups, a separate input matrix is analysed allowing for the evaluation of measurement invariance (e.g., equality of indicator intercepts) and population heterogeneity (e.g., differences in latent means) to be assessed by constraining like parameters in each of the groups (such as factor loadings; Brown, 2006).

It is acknowledged that within the literature pertaining to CFA methodologies, there is some variability in the terminology used. For example, Brown (2006) notes that the test of equality of factor structures (indicating that the number of factors and patterns of indicator-factor loadings is equivalent across groups) has also been referred to as “configural invariance”. “Metric invariance” and “weak factorial invariance” have been used to refer to the equality of factor loadings. “Scalar invariance” and “strong factorial invariance” have been used as alternatives to the equality of indicator intercepts. Finally, a test of “strict factorial invariance” has been used to refer to as an evaluation of the equality of indicator residuals (Meredith, 1993). Brown (2006) indicates that “a more descriptive and pedagogically useful terminology is encouraged...” (p.268). For this reason, following Brown’s recommendations, the terms *equal form, equal factor loadings,* and *equal intercepts* will be used in the description of measurement properties in the coming chapters.

Additionally, there is some disagreement as to the order in which the model restrictions are evaluated (Brown, 2006; Horn & McArdle, 1992). A step-wise approach is recommended in order to identify multiple sources of non-invariance. This is considered a more prudent approach as it is possible to determine if further test of invariance and homogeneity are required. Additionally, latent means group differences are considered meaningful only if the factor loadings and the indicator intercepts are found to be invariant, while comparisons of factor variances and covariances are only meaningful if the factor loadings and indicator intercepts are found to be invariant.

Brown (2006) recommends the following progressive analytic strategy for the assessment of measurement invariance: (1) verify the CFA model for each group separately; (2) conduct simultaneous tests of equal factor structure (i.e., equal form); (3) verify the equality of factor loadings; (4) determine the equality of indicator intercepts; and (5) test the equality of residual variances of the indicators. This progressive approach is particularly relevant when some non-variant parameters (i.e., unequal factor loadings) are encountered. It is possible to proceed within the context of *partial measurement invariance* where some of the measurement parameters are not equal (Brown, 2006; Byrne, Shavelson, & Muthén, 1989). For the assessment of population heterogeneity, the following is recommended: (6) test the factor variances for equality; (7) assess the equality of factor covariances (if there is more than one factor in the model); and (8) verify the equality of latent means. It is noted that the assessment of factor covariances is not relevant in the current dissertation as the MG-CFA measurement model comprises a one-factor representation of general intelligence (or *g*).

4.5.1. Model Specificity and Evaluation

It is acknowledged that IRT and SEM techniques offer several advantages to Classical Testing methods. However, unless the model used for parameter estimation sufficiently fits the data, the benefits of the methodology may not be realised. As such, the choice of model parameters must be made on both theoretical and empirical grounds (Baker, 2001).

Within the common factor model there are parameters that are known and those that are unknown. The 'known' parameters of the model are the indicator variance and covariances; these provide information needed for a model and are often referred to as the 'variance-covariance matrix' (Brown, 2006). For categorical data, as in the case of the SPM+, a correlation matrix is used in its place. The 'unknown' parameters of the model are the factor loadings, error variances of the indicators, and factor variances. Factor loadings are the regression slopes used to predict indicators from latent variables. Error variances are a reflection of the remaining variance that cannot be accounted for by the latent construct. Factor variances reflect differences within the sample on a latent construct, including sampling error.

In situations where multiple indicators demonstrate similarity that is not explained by the latent trait, it may be necessary to specify error covariances, or correlations between the indicators and the error variances of the indicators. In models that specify more than one latent construct, it is possible to specify factor covariances which are correlations between similar latent constructs (Brown, 2006). Each of the unknown parameters is estimated by the software in order to replicate the variance-covariance matrix (or the correlation matrix for categorical data) as closely as possible.

The software used in this dissertation, Mplus version 5.21 (Muthén & Muthén, 2009) uses a *fitting function* to minimise the difference between the observed and predicted matrices. The fitting function most widely used, and employed in the MG-CFA models of this dissertation is the *maximum likelihood* (ML) function. The objective of ML estimation is to find parameter estimates that most closely resemble the data if the data were collected from the same population again and to minimise the differences between the variance-covariance or correlation matrices (Brown, 2006).

In order to identify the model, the software must first assign a scale to the latent variable; as it is an unobserved construct the latent variable has no defined metric. The most common approach to defining the metric of the latent variable, as is the method employed in this thesis, is to scale the latent variable according to the first indicator (also referred to as the *marker* or *reference indicator*). Next, the model must ensure that the number of freely estimated (or unknown) parameters does not exceed the number of known parameters. If this is achieved, the model is said to be 'over-identified': the model will converge and model fit statistics will be provided (discussed below).

Convergence of a model is reached when the software programme reaches a set of parameter estimates that cannot be improved upon further to reduce the difference between the observed and estimated variance-covariance matrices. The ability of a model to converge successfully is related to the quality and complexity of the specified model (Brown, 2006). A model is said to be 'just-identified' if the number of known parameters is equivalent to the number of unknown parameters: the model will still converge but model fit statistics will not be computed. In cases where the number of unknown parameters exceeds the number of known parameters, the model is said to be under-identified and will not converge. To avoid such problems, three indicators per latent variable are recommended (Brown, 2006).

Once the model has converged successfully, it is necessary to evaluate whether the estimated model is a good representation of the observed data. Model fit can be assessed in three different ways: absolute fit, fit adjusting for model parsimony, and incremental or comparative fit.

4.5.1.1. Absolute Fit

Absolute fit indices assess how well the model fits the hypothesis that the observed and estimated variance-covariance matrices are equivalent. The chi-square statistic can be considered a classic method of assessing overall model fit. The software calculates the difference between the observed and model-estimated associations between the indicators in the variance-covariance matrices. If the difference is found to be significant, the model is not considered to have fit the data very well.

However, the chi-square statistic is sensitive to a number of factors that detract from its reliability as a sole estimator of global model fit. The chi-square statistic can be artificially inflated by a large sample size, and it is considered an overly stringent test due to the hypothesis that the observed and expected values are identical (Brown, 2006). It is therefore suggested that the chi-square be interpreted in conjunction with other fit indices, such as the *standardised root mean square residual* (SRMR) or the *weighted root mean square residual* (WRMR).

The SRMR and the WRMR are statistics of the average discrepancy between correlations observed in the input matrix and the correlations predicted by the model, and range in value between 0.0 and 1.0, where the smaller values indicate a better fit of the model (Brown, 2006). It is noted that SRMR is not suitable for use with binary data (Yu, 2002), and the WRMR is not suitable for use with multiple group models (Brown, 2006). Therefore, the chi-square statistic will be evaluated in this dissertation in conjunction with parsimony correction and comparative fit indices.

4.5.1.2. Parsimony Correction

Indices of parsimony correction are similar to the indices of absolute fit but they account for how precise and parsimonious the model is. For example, consider a situation where two different models, Model 1 and Model 2, fit a sample matrix equally well in terms of absolute fit. However, Model 2 has more freely estimated parameters (and hence, more degrees of freedom) than does Model 1. According to indices of parsimony correction, Model 1 would be preferred because the model fits the sample data with the fewest freely estimated parameters, and is therefore considered to be a more parsimonious solution.

The recommended index for parsimony correction is the *root mean square error of approximation* (RMSEA; Steiger & Lind, 1980). The RMSEA is based upon a non-central chi-square distribution, and is thus considered to be an 'error of approximation' index. It assesses the extent to which the model fits reasonably well in the sample population, as opposed to assessing whether the model fits the sample population exactly as in the case of the chi-square statistic. As with the SRMR, RMSEA values of 0 (or very close to 0) indicate a perfect model fit. At its upper range, the RMSEA is unbounded but it is rare to see it exceed 1.0 (Brown, 2006).

4.5.1.3. Comparative Fit

Comparative fit indices are so named because the statistic evaluates the fit of the model in relation to a more restricted baseline model. Generally, this baseline model is a 'null' model where the covariances among the indicators are set to zero. One of these indices the *comparative fit index* (CFI; Bentler, 1990) is based on the non-centrality parameter (like the RMSEA), and uses information from expected values of the chi-square statistic. The CFI has a range of values of 0.0 to 1.0, with values nearer to 1.0 indicating good model fit.

Another widely-used index of comparative fit is the Tucker-Lewis Index (TLI; Tucker & Lewis, 1973). The TLI assesses model complexity in relation to the baseline model (as does the RMSEA) and includes a penalty function for the addition of freely estimate parameters that do not sufficiently improve the fit of the model. Unlike the CFI, the TLI is non-normed; the values of TLI can fall outside the range of 0.0 to 1.0. However, like the CFI, values of TLI approaching 1.0 indicate good model fit (Brown, 2006).

4.5.1.4. Interpretation of Goodness-of-Fit

There is much debate surrounding the interpretation of goodness-of-fit indices, and what are deemed suitable recommended index cut-off ranges. It is important to note that goodness-of-fit indices are but one aspect of model evaluation. It is equally important to determine whether there are any points of localised strain in the model, as well as the interpretability and strength of the resulting parameter estimates of the model (Brown, 2006).

Two parameters are of particular interest in model fit assessment: (1) factor loadings and (2) the amount of indicator variance that is explained by the latent variables, or the explained variance divided by the total variance (R^2). Both the factor loadings and the R^2 values of the model should be statistically significant and salient (approximately 0.30 or greater; Brown, 2006). If this is achieved, the observed measures can be considered reliable indicators of the underlying construct, and it is suitable to proceed with the assessment of goodness-of-fit indices.

According to Brown (2006) and Hu and Bentler (1999), the following are suggested guidelines for the evaluation of model goodness-of-fit: (1) a significant chi-square statistic; (2) RMSEA $\leq$ 0.06; and (3) CFI and TLI $\geq$ 0.95. These values are not absolute, and ought to be considered only as guidelines. Values of these indices have been found to fluctuate as a function of modelling conditions, and thus values that are slightly out of the ranges indicated above can still be considered acceptable (Brown, 2006; Hu & Bentler, 1999).

For example, values of RMSEA between 0.05 and 0.08 suggest adequate model fit, values less than .05 suggest good model fit, and that models $\geq$ 0.10 ought to be rejected (Browne & Cudek, 1993; MacCallum, Browne, & Sugawara, 1996). Similarly, CFI and TLI indices with values in the range of 0.90-0.95 may indicate acceptable model fit, while values below .90 indicates unacceptable fit that is often worthy of model rejection (Bentler, 1990). It is further emphasised that goodness-of-fit indices not be used stringently and exclusively in the assessment of model fit. Rather, it is important to consider model fit in light of a number of different fit indices in tandem, as well as accounting for theoretical and practical aspects of the analytic situation.

While the goodness-of fit statistics (chi-square, RMSEA, CFI, TLI) provide an overall, global indication of how well the model reproduces the observed relationships among indicators in the input matrix, they do not provide an assessment of how well the model replicates indicator-level relationships. Modification indices provide an approximation of how much the chi-square statistic (with 1 degree of freedom) would decrease if a constrained or fixed parameter were allowed to be freely estimated (Brown, 2006). Modification indices of 3.84 or greater (which reflect the chi-square critical value at $p < 0.05$ and 1 degree of freedom) suggest that the model could be respecified to significantly improve model fit.

4.5.1.5. Model Respecification

Once the model has been assessed for specificity and goodness-of-fit according to the guidelines indicated above, it may be necessary to respecify the model in order to improve the fit of the model if the minimum criteria of fit have not been realised. The main sources of model misspecification are the number of factors (either too few or too many), the indicators or pattern of indicator-factor loadings, and the management of measurement error (e.g., uncorrelated vs. correlated measurement error). It is not appropriate to respecify the model in response to modification indices with the sole intention of improving model fit. Rather, model respecification should be guided by compelling substantive evidence and only in accordance with empirical, practical or conceptual grounds.

4.6. Summary

It has been argued that classical test theory methods are insufficient for the assessment of group differences in terms of both strength and specificity. A more robust and reliable alternative of investigating relationships between observed variables and latent factors using structural equation modelling has been described in this chapter.

First, item response theory (IRT) was introduced. Classical methods and IRT are similar in the conceptualisation of individual item responses as they relates to the underlying latent construct. However, IRT is considered superior in its ability to model the variation of the likelihood of endorsing an item as a function of the respondent's underlying level of the latent construct. IRT methods will be employed in Chapter 5 to evaluate item level characteristics of the SPM+.

Moving from item-level analysis to assessing the latent construct in the overall test, factor analysis methods were described. When considering a collection of related items on a test, factor analysis techniques provide the tools to help understand the variation and covariation amongst items in the most parsimonious manner.

Exploratory factor analysis is most usefully employed in the early stages of analysis to ascertain the number of underlying factors exist within the data, while CFA offers the researcher the ability to verify solutions suggested by EFA as well as provide greater control over the model specification (Brown, 2006). This allows CFA techniques to effectively address questions relating to group differences, such as sex differences on the Raven's SPM+.

In the following results chapters, the reader will notice a step-wise, cumulative progression to the strategy of analyses. Chapter 5 begins the investigation of the Raven's SPM+ from the item-level perspective using Item Response theory. Through the use of a Rasch model and item characteristic curves, the characteristics of the items are evaluated as they relate to the underlying latent construct of general intelligence.

The investigation of group-level differences begins in Chapter 6 using Multiple-Groups Confirmatory Factor Analysis (MG-CFA) to assess all aspects of measurement invariance and population homogeneity for males and females – factor loadings, indicator intercepts, residual variances, factor variances, and latent means. The effects of age will be introduced into the analyses in Chapter 7. The literature surrounding the "developmental theory of sex differences" (Lynn, 1999) indicates that girls mature earlier than boys, both cognitively and physically, resulting in a female cognitive advantage over males until approximately 15 years of age, at which point males outperform females (Lynn, 2002; Lynn, Allik, & Must, 2000).

In order to address the question of sex differences at different points along the developmental continuum, MG-CFA analyses will be conducted to assess latent mean differences of boys and girls in two different age groups: 7-14, and 15-18 years. In both Chapters 6 and 7 the unidimensional nature of the SPM+ data will be tested further through the addition of methods factors to the models.

— 5 —

Results Chapter 1:

the Raven's SPM+ at the Item Level - Item Response Theory & The Rasch Model

5.1. Introduction

To begin answering the main aims of this dissertation in a psychometrically appropriate manner, it is important to first begin to understand what the Raven's Standard Progressive Matrices (SPM) is measuring at the item level. The SPM is a measure commonly used in investigations of sex differences in general intelligence, or *g* (Raven, Court, & Raven, 2008; Lynn, Allik, & Irwing, 2004). The SPM was constructed as a measure of the educative component of *g*, which is the ability to forge new insights, to discern meaning in confusion, to perceive, and to identify relationships (Spearman, 1927). Despite the measure's specific design to measure one latent factor of general intelligence, the underlying factor structure of this non-verbal measure of abstract reasoning has been debated.

As discussed in Chapter 2, some researchers claim that the Raven's Matrices assess multiple factors of intelligence. Lynn, Allik, and Irwing, (2004) argued for a 3-factor structure where items loaded onto Gestalt continuation, Verbal-analytic reasoning, and Visuospatial ability factors. Others, such as van der Ven and Ellis (2000) posit a 2-factor structure where the SPM measures 'g' and a second perceptual or spatial factor.

In contrast, many believe the Raven's to be among the best measures of unidimensional construct of general intelligence (Abad, Colom, Rebollo, & Escorial, 2004; Court, 1983; Jensen, 1998; Raven, 2009), and to have "the highest *g* loading" of any measure of cognitive ability (p.541, Jensen, 1998). By applying a one-factor model, the implication is that the Raven's measures a single, underlying latent trait of general intelligence.

Gaining a better understanding of the factor structure and, ultimately, what the Raven's SPM+ is measuring is vital to properly answering the overall research aims of this dissertation and the question of sex differences in general intelligence. Most importantly, if the Raven's Matrices provides a pure measure of g, any mean score differences or differences in variability on Raven's performance may lead to conclusions relating to sex differences in general intellectual functioning.

5.2. Factor Analysis

As there is an established literature debating the factor structure of the Raven's Matrices, it was considered important in the current dissertation to verify the disparate claims. Establishing a suitable factor structure begins with factor analysis. The fundamental objective of factor analysis is to determine the number and nature of the latent constructs, or 'factors', that account for the variation and covariation among a set of test items (Brown, 2006). A factor can be considered an unobservable construct that is common among a set of test items, and helps to explain how a sub-set of items are correlated. There are two forms of factor analysis, exploratory and confirmatory, both of which are used in this chapter.

5.2.1. Exploratory Factor Analysis

As detailed in Chapter 4, the main objective of an exploratory factor analysis (EFA) is to evaluate the dimensionality of a set of test items (i.e., indicators) by determining the fewest number of interpretable factors needed to explain the covariation among items. In this dissertation, EFA will be used to verify claims previously made with respect to the unidimensional versus multidimensional nature of the Raven's Matrices.

As previously noted, EFA is fundamentally distinct from its counterpart Principal Components Analysis (PCA), which often is mistakenly used in the place of EFA. The primary goal of PCA is data reduction – to reduce a large number of indicators to a smaller set of variables that account for the large amount of observed variance (Kashy et al., 2009). PCA is not considered robust or sensitive enough to fully account for the underlying construct in the current measurement model because it does not differentiate between common and unique variance. Therefore, EFA techniques were chosen for this dissertation.

EFA is exploratory in nature because no *a priori* restrictions are imposed upon the model, and is often conducted as a precursor to conducting a Confirmatory Factor Analysis (CFA). As detailed in section 4.3.1, there are five main steps of an EFA. First, the researcher must determine whether EFA techniques are appropriate for the empirical objectives, the size and nature of the sample, and if so, which of the test indicators to include in the analyses. Next, factors are extracted, followed by factor selection. As indicated in Brown, 2006, the determination of the number of factors is considered crucial because “under-factoring” (the selection of too few factors) or “over-factoring” (the selection of too many factors) can seriously compromise the validity of the resulting factor model. In this chapter, two procedures of factor selection based on eigenvalues were employed: the *Kaiser-Guttman* rule, and examination of scree plots. Eigenvalues can be viewed as representing the variance in the indicators explained by the successive factors.

Finally, the extracted factors are rotated, using an oblique method, to increase their interpretability. An oblique method of rotation was used in this dissertation, as it provides a more realistic representation of how factors are inter-correlated. Further, if the EFA is used as a precursor to CFA, as is the case in this dissertation, oblique rotations are more likely to generalise to CFA solutions than orthogonal rotation methods (Brown, 2006; Kashy et al., 2009).

Although the process of EFA is an exercise to explore and describe the underlying factor structure of the test, the decision of the appropriate number of factors should be guided by substantive considerations as well as statistical guidelines (Brown, 2006; Costello & Osborne, 2005). These will be discussed further. Upon the determination of a suitable factor structure, the SPM+ data is modelled using an Item Response Theory Rasch model, also known as a one-parameter logistic confirmatory factor analysis model.

5.3. The Rasch Model

Item Response Theory allows for a better understanding of the relationship between item characteristics (i.e., item parameters) of the SPM+ and innate characteristics of the test respondents (i.e., latent traits) of the standardisation sample, and how these relate to the likelihood of endorsing a particular response category (Brown, 2006). Results of the IRT can be evaluated according to three different item parameters: discrimination, difficulty and a pseudo-guessing parameter. The discrimination of an item describes how well an item can distinguish between participants of different levels of the latent trait of cognitive ability. The difficulty of an item corresponds to the value of the latent trait at the point where the predicted probability of a respondent correctly endorsing an item is 50%. The pseudo-guessing parameter relates to the element of random guessing used by test respondents, regardless of the level of the latent trait. The random guessing strategy enables the respondent to select the correct answer with a probability of 1/β, where β is the number of possible response categories in an item with multiple response options. Because IRT can be considered a form of regression procedure, the item and participant parameters do not vary in accordance with ability level. It is therefore possible to determine the contribution and characteristics of each item independently (Hulin et al., 1983).

Within the context of determining the factor structure of the SPM+, a CFA is used to verify the factor structure proposed by the EFA, and allows the strength of relationship to be tested between the observed variables (or indicators) and their underlying latent constructs. Every aspect of the model is pre-specified by the researcher and the acceptability of the model is evaluated using goodness-of-fit and interpretability of the parameter estimates (as discussed in Chapter 4).

A Rasch model is a one-parameter logistic IRT model (a form of CFA) where the discrimination parameter of each item in the model is held to equality in order to ascertain the difficulty of each item. In reference to Equation 2 described in section 4.2, the Rasch model is obtained by modifying the normal ogive curve by setting the guessing parameter to zero (c = 0), constraining the discrimination of all items to equality (a = 1) and allowing the difficulty parameter (b) to be freely estimated. In other words, the model is allowing the difficulty of each item to be freely estimated, while imposing a restriction so that the likelihood of answering an item correctly is equivalent across all levels of ability level.

Recalling the shape of the normal ogive curve from Chapter 4, the discrimination of an item influences the slope of the curve. By constraining the discrimination to equality in the Rasch model, all items display equal slopes, thereby illustrating the difficulties of the items along the horizontal axis of the Item Characteristic Curves (ICCs). For ease of interpretation, item curves are displayed in this chapter in item sets for males and females separately allowing for a comparison of items, and ultimately, allowing for the verification of item difficulty within and across item sets.

Overall, the aim of this chapter is to provide a thorough assessment of the dimensionality of the Raven's Standard Progressive Matrices Plus (SPM+) in order to establish a suitable model upon which subsequent analyses will be based.

5.4. Understanding the SPM+ at the Item Level

5.4.1. Descriptive Information

In order to better understand the SPM+ as a whole, it is important to fully understand it at the item-level. As detailed in Figure 5, items on the SPM+ were scored as either correct or incorrect, and were entered into the response database as binary values of the indicators. Item means and standard deviations for males and females in the overall sample are provided in Table 5.

Table 5: Item means and standard deviations for males and females

SPM+ Items	Male		Female		SPM+ Items	Male		Female	
	Mean	SD	Mean	SD		Mean	SD	Mean	SD
A3	0.99	0.12	0.98	0.13	**C8**	0.50	0.50	0.60	0.49
A4	0.98	0.15	0.99	0.11	**C9**	0.22	0.41	0.21	0.41
A5	0.97	0.16	0.98	0.15	**C10**	0.11	0.31	0.08	0.27
A6	0.98	0.15	0.98	0.13	**C11**	0.10	0.30	0.12	0.32
A7	0.90	0.30	0.91	0.28	**C12**	0.15	0.36	0.19	0.40
A8	0.86	0.34	0.85	0.36	**D1**	0.60	0.49	0.63	0.48
A9	0.93	0.26	0.93	0.26	**D2**	0.51	0.50	0.52	0.50
A10	0.89	0.31	0.84	0.37	**D3**	0.46	0.50	0.46	0.50
A11	0.76	0.43	0.74	0.44	**D4**	0.22	0.41	0.24	0.43
A12	0.63	0.48	0.64	0.48	**D5**	0.16	0.37	0.16	0.37
B1	0.97	0.18	0.97	0.16	**D6**	0.20	0.40	0.25	0.43
B2	0.97	0.16	0.97	0.18	**D7**	0.29	0.45	0.23	0.42
B3	0.96	0.19	0.98	0.15	**D8**	0.14	0.35	0.14	0.35
B4	0.87	0.33	0.87	0.33	**D9**	0.22	0.42	0.26	0.44
B5	0.93	0.26	0.90	0.31	**D10**	0.16	0.37	0.12	0.33
B6	0.80	0.40	0.78	0.42	**D11**	0.13	0.33	0.15	0.36
B7	0.74	0.44	0.78	0.42	**D12**	0.08	0.28	0.11	0.32
B8	0.79	0.40	0.83	0.37	**E1**	0.38	0.49	0.39	0.49
B9	0.81	0.39	0.85	0.36	**E2**	0.18	0.39	0.24	0.43
B10	0.86	0.34	0.91	0.29	**E3**	0.29	0.45	0.30	0.46
B11	0.75	0.43	0.83	0.38	**E4**	0.21	0.40	0.18	0.39
B12	0.64	0.48	0.65	0.48	**E5**	0.08	0.27	0.07	0.25
C1	0.90	0.30	0.90	0.30	**E6**	0.14	0.35	0.15	0.35
C2	0.90	0.30	0.92	0.28	**E7**	0.10	0.30	0.14	0.35
C3	0.75	0.43	0.75	0.43	**E8**	0.08	0.26	0.07	0.26
C4	0.93	0.26	0.94	0.24	**E9**	0.05	0.23	0.10	0.30
C5	0.62	0.49	0.58	0.49	**E10**	0.09	0.28	0.08	0.27
C6	0.80	0.40	0.84	0.36	**E11**	0.08	0.28	0.07	0.25
C7	0.40	0.49	0.37	0.48	**E12**	0.05	0.23	0.07	0.26

Total scores on the SPM+ are obtained by summing the correct scores for each individual, with the highest possible total score of 59. Mean, minimum and maximum total scores are available for males and females for the overall sample, and for the younger and the older age samples in Table 6. The range of total scores for the U.K. standardisation sample was 8-53 (mean = 31.27, SD = 7.086) for males, and 7-50 (mean = 31.76, SD = 6.954) for females.

Table 6: Mean, minimum and maximum total scores for the overall, younger and older samples of the SPM+

		Overall (7-18 years)	**Younger (7-14 years)**	**Older (15-18 years)**
Male	**N**	437	323	114
	Min	8	8	21
	Max	53	50	53
	Mean	31.27	29.75	35.60
	SD	7.09	6.76	6.15
Female	**N**	489	340	149
	Min	7	7	18
	Max	50	46	50
	Mean	31.76	29.69	36.46
	SD	6.95	6.57	5.34

5.5. Exploratory Factor Analysis and the SPM+

Exploratory factor analysis techniques were deemed suitable for use with the SPM+ due to the magnitude of the sample size (N = 926) and the magnitude of the subject to item ratio (16:1; Costello & Osborne, 2005; Guadagnoli & Velicer, 1988). In order to fully understand the factor structure of the SPM+ in the format in which it is most commonly used in research and private practice, 58 of the 60 test items were retained in the exploratory analyses. For practical reasons, two items needed to be removed from the analyses. Item A1 is provided to all participants as a training item and is, therefore, not included in any of the analyses in this dissertation. Item A2 is the first test item on the measure, and all participants in the sample achieved a correct score on this item. It was therefore not included in the analyses as it did not meaningfully contribute to understanding the relationship between the content of the SPM+ and the latent construct of general intelligence.

An exploratory factor analysis with the SPM+ data was conducted in Mplus Version 5.21 (Muthén & Muthén, 2009). The dimensionality of the 58 test items was extracted using robust weighted least means squares (WLSMV) estimation. WLSMV estimation was chosen because of the categorical nature of the data. Further, it is a full information estimator that provides an estimation of how well the factor solution is able to reproduce the relationship among indicators in the correlation matrices (Brown, 2006).

Two procedures of factor selection based on eigenvalues were employed: the Kaiser-Guttman rule, and examination of scree plots. Eigenvalues for the sample correlation matrix are provided for the overall sample, as well as for females and males separately in Table 7. The Kaiser-Guttman rule identified 19 suitable factors for the overall sample, 20 factors for the female sample and 21 factors for the male factors greater than the minimum threshold of 1 eigenvalue.

Next, scree plots were consulted for the male and female samples (Figure 11 and Figure 12 respectively). The plots were inspected to determine the point at which the last significant decline in magnitude of eigenvalues is located. Upon inspection, the scree plots for males and females demonstrate there are two points of significant decline of eigenvalues: between factors two and three, and between three and four. The same pattern of eigenvalue decline was found by Lynn et al. (2004).

Table 7: Eigenvalues from the EFA from the Female and Males samples of the SPM+

Component	Eigenvalues for Females	Eigenvalues for Males
1	16.065	15.817
2	4.533	5.528
3	2.813	3.122
4	2.537	2.555
5	2.427	2.475
6	2.310	2.110
7	2.035	2.040
8	1.775	1.913
9	1.731	1.805
10	1.704	1.691
11	1.666	1.649
12	1.661	1.631
13	1.473	1.592
14	1.398	1.463
15	1.356	1.359
16	1.327	1.330
17	1.225	1.241
18	1.174	1.200
19	1.052	1.124
20	1.023	1.082
21	0.983	1.010
22	0.912	0.978

Figure 11: Scree plot of eigenvalues for the male sample of the SPM+

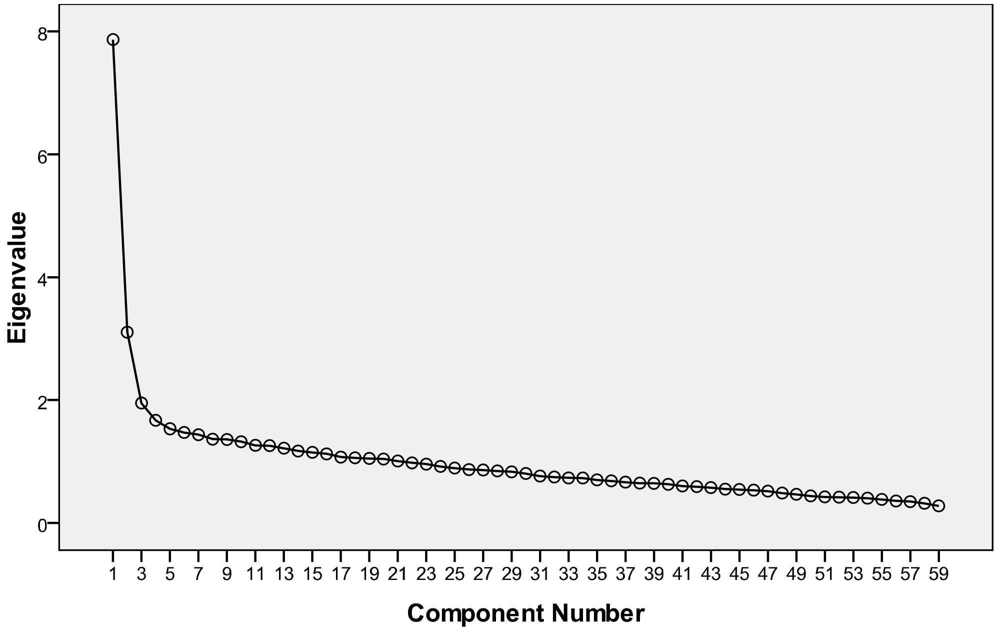
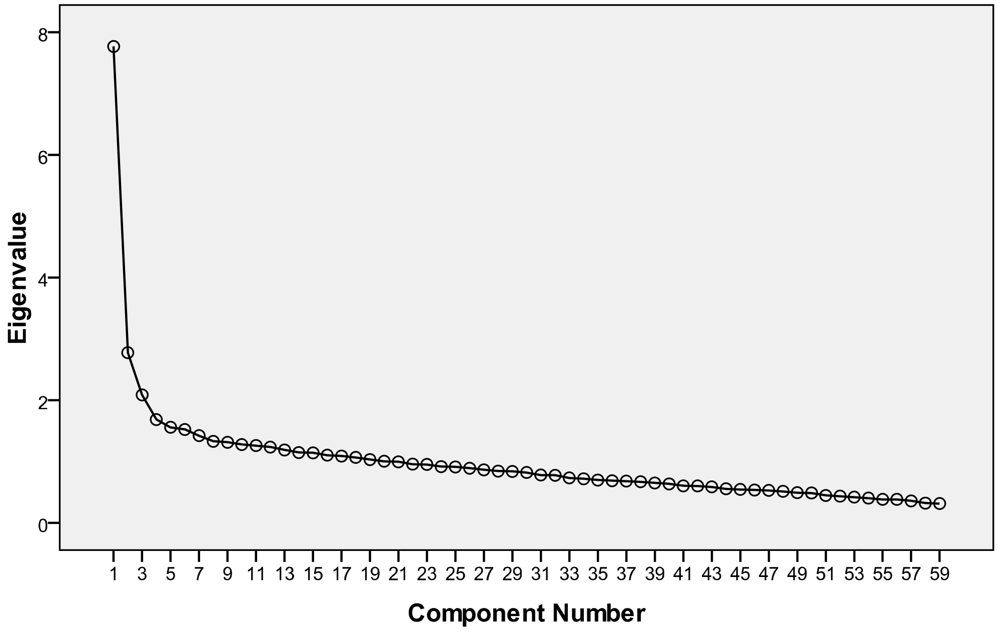


Figure 12: Scree plot of eigenvalues for the female sample of the SPM+



One noted limitation of this method is that the interpretation of the point where the magnitude of eigenvalues change is subjective and open to interpretation. It is for this reason that multiple methods of eigenvalue evaluation are used.

In addition to the eigenvalue-based procedures for factor estimation (i.e., the Kaiser-Guttman rule and the scree test), WLSMV estimation has the advantage of being a full-information estimator that provides goodness-of-fit indices of the model that can be used to determine the appropriate number of factors. Table 8 provides the goodness-of-fit statistics for the one-, two-, and three-factor solutions that were indicated by the scree plots.

Table 8: Goodness-of-fit statistics for one, two-, and three-factor solutions for the SPM+

	# of factors	**χ^2 (*df*)**	***p***	**CFI**	**TLI**	**RMSEA**	**SRMR**
Overall	1	3585.342 (1595)	0.001	0.928	0.925	0.037	0.103
	2	2728.283 (1538)	0.001	0.957	0.953	0.029	0.089
	3	2083.374 (1482)	0.001	0.978	0.976	0.021	0.081
Males	1	2895.325 (1652)	0.001	0.916	0.913	0.041	0.142
	2	2271.613 (1594)	0.001	0.954	0.951	0.031	0.121
	3	1862.210 (1537)	0.001	0.978	0.976	0.022	0.110
Females	1	2752.780 (1652)	0.001	0.929	0.926	0.037	0.126
	2	2306.222 (1594)	0.001	0.954	0.951	0.030	0.110
	3	2053.353 (1537)	0.001	0.967	0.963	0.026	0.104

In each of the factor solutions, the models for the overall sample were over-identified, with chi-square values ranging from 2083.374 (*df* = 1482) to 3585.342 (*df* = 1595), and showed good model fit with the following ranges of values: CFI = 0.928 - 0.978; TLI = 0.925 – 0.976; RMSEA = 0.021 - 0.037; and SRMR = 0.081 - 0.103. For males, the models for the overall sample were also over-identified, with chi-square values ranging from 1862.210 (*df* = 1537) to 2895.325 (*df* = 1652), and showed good model fit with the following ranges of values: CFI = 0.916 - 0.978; TLI = 0.913 – 0.976; RMSEA = 0.022 - 0.041; and SRMR = 0.110 – 0.142. For females, the models for the overall sample were over-identified, with chi-square values ranging from 2053.353 (*df* = 1537) to 2752.780 (*df* = 1652), and showed good model fit with the following ranges of values: CFI = 0.929 - 0.967; TLI = 0.926 – 0.963; RMSEA = 0.026 - 0.037; and SRMR = 0.104 – 0.126.

Next, the extracted factors were rotated using an oblique method to increase their interpretability. Details of the minimum, maximum, and average factor loadings for the one-, two-, and three-factor solutions are presented in Table 9. Upon further inspection of the factor loadings, a number of 'poorly behaved items' (Brown, 2006) were identified. There were a number of items that showed low communalities, or factor loadings which fall below 0.3 (McDonald, 1999) suggesting that the indicator is not meaningfully representative of the factor upon which it is loaded. Further, a number of the indicators loaded upon more than one factor simultaneously, also known as cross-loadings.

It is advised (McDonald, 1999; Brown, 2006) that indicators that do not reach the minimum factor loading threshold of 0.3, or load upon multiple factors simultaneously should not be included in the analyses, unless there are theoretical grounds to do so. Because the SPM+ is a published test of intelligence that is used extensively in research and practice, all test items will be retained in further analyses for reasons of completeness, comparability to other published findings, and for generalisability to the population at large.

Table 9: Minimum, maximum, and average factor loadings for the one-, two-, and three-factor solutions

		Factor Loadings			
		Minimum	Maximum	Average	# of non-sign or non-salient F.L.
1-factor	Overall	-0.021	0.843	0.479	10
	Male	-0.043	0.883	0.480	10
	Female	-0.055	0.831	0.476	12
2-factor	Overall	F1: -0.263 F2: -0.405	F1: 0.921 F2: 0.718	F1: 0.343 F2: 0.191	26 35
	Male	F1: -0.246 F2: -0.148	F1: 0.968 F2: 0.726	F1: 0.308 F2: 0.248	28 29
	Female	F1: -0.569 F2: -0.264	F1: 0.870 F2: 0.693	F1: 0.261 F2: 0.289	33 23
3-factor	Overall	F1: -0.133 F2: -0.324 F3: -0.267	F1: 0.902 F2: 0.801 F3: 0.551	F1: 0.214 F2: 0.302 F3: 0.136	35 24 38
	Male	F1: -0.369 F2: -0.409 F3: -0.257	F1: 0.884 F2: 0.757 F3: 0.675	F1: 0.243 F2: 0.178 F3: 0.214	34 34 31
	Female	F1: -0.277 F2: -0.451 F3: -0.263	F1: 0.896 F2: 0.816 F3: 0.601	F1: 0.203 F2: 0.252 F3: 0.173	39 29 37

The selection of the appropriate number of factors ought to take place within the context of substantive considerations, not statistical guidelines alone (Brown, 2006). While the existing literature pertaining to the factor structure of the Raven's is inconsistent, there is sufficient support for a one-factor structure (Raven, 2009; Silverman et al., 2000). The current EFA results offer sufficient support for a one-factor structure, and the suitability for use with the U.K. standardisation data will therefore be tested further. A

Rasch model will now be used to assess the suitability of a one-factor model of the SPM+ at the item level.

5.6. Rasch Model

In the Rasch model (also known as the *one-parameter logistic* model), the discrimination parameter of each item in the model is held to equality in order to ascertain the difficulty of each item. The 58 test items of the SPM+ were loaded onto one latent factor and analysed using Maximum Likelihood estimation and Rasch scaling. Maximum Likelihood (ML) is a full information estimator allowing for an assessment of how well the model is able to reproduce the observed variances and covariance among the input indicators.

In order to better understand item-level performance for males and females on the SPM+, analyses were conducted for the two samples separately. For males, the Rasch model (Model M5-1a) was over-identified with 120 df; $\chi^2 = 436.875$, $p < 0.001$. The RMSEA indicates that the model adequately fits the data: RMSEA = 0.078; CFI = 0.676; TLI = 0.695. For females, the Rasch model (Model M5-1b) was over-identified with 132 df; $\chi^2 = 523.571$, $p < 0.001$. As with males, the RMSEA provides sufficient evidence that the model adequately fits the data: RMSEA = 0.078; CFI = 0.639; TLI = 0.672. Details of item difficulties and average item difficulty are presented by item sets in Table 10 through Table 16 in the following section.

In addition to the assessment of model fit, it remains important to review graphical representations of the items of the SPM+ in order to better understand the performance of males and females of the standardisation sample. Item characteristic curves (ICCs) illustrate a non-linear regression of the probability of obtaining a correct response along the continuum of variation in the latent variable of intellectual ability (Zumbo, 1999). For ease of interpretation, the ICCs for males and females have been grouped into item sets A-E, Figure 13 to Figure 22 respectively.

The SPM+ was originally designed such that items increase in a step-wise progression of difficulty, with the easiest items at the beginning of the item set, becoming increasingly difficult with each subsequent item and with each subsequent item set. The ICCs will now be reviewed with the intention of confirming the increase of item difficulties within and across item sets.

5.6.1. Set A

Item difficulty values for items A3 to A12 of set A of the SPM+ are provided in Table 10. These item difficulties are graphically presented in Figure 13 for females and in Figure 14 for males.

Table 10: Item difficulties of Items in Set A for Males and Females

	Male				Female			
SPM+ Items	**Est.**	**S.E.**	**Est./S.E.**	**2-tailed p-value**	**Est.**	**S.E.**	**Est./S.E.**	**2-tailed p-value**
A3	-3.539	0.281	-12.609	0.00	-3.622	0.424	-8.533	0.00
A4	-2.919	0.240	-12.138	0.00	-2.978	0.361	-8.254	0.00
A5	-2.828	0.227	-12.482	0.00	-2.884	0.353	-8.170	0.00
A6	-2.911	0.236	-12.343	0.00	-2.971	0.356	-8.333	0.00
A7	-2.005	0.161	-12.488	0.00	-2.030	0.159	-12.754	0.00
A8	-1.511	0.121	-12.48	0.00	-1.518	0.126	-12.044	0.00
A9	-2.250	0.176	-12.809	0.00	-2.285	0.229	-9.973	0.00
A10	-1.672	0.136	-12.339	0.00	-1.685	0.146	-11.538	0.00
A11	-1.049	0.100	-10.458	0.00	-1.038	0.112	-9.272	0.00
A12	-0.579	0.090	-6.448	0.00	-0.550	0.128	-4.306	0.00
Average	-2.126				-2.156			

Figure 13: Female Sample Item Characteristic Curves for the Rasch Model of Item Set A of the Raven's SPM+

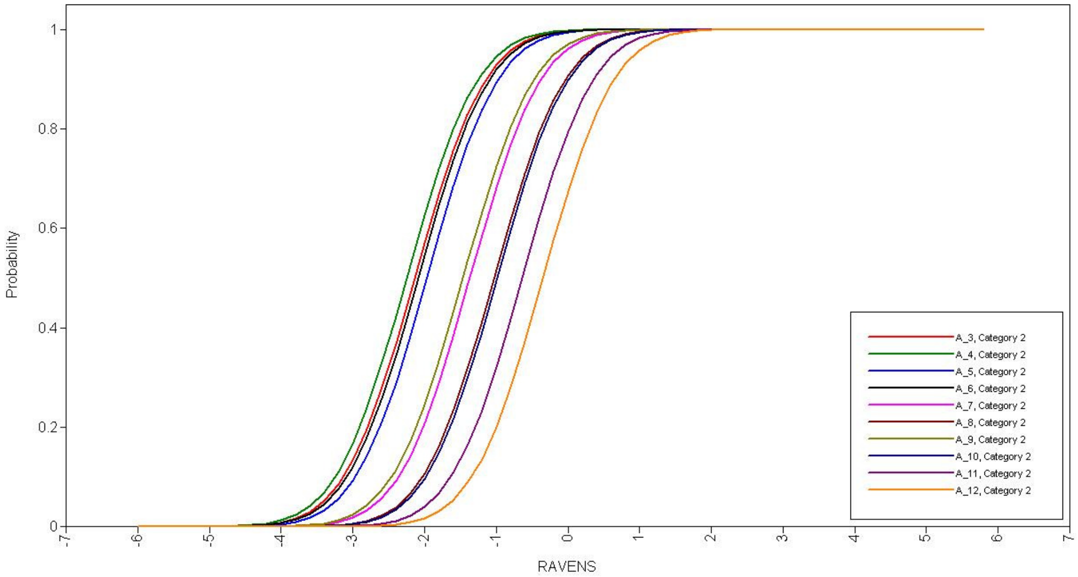
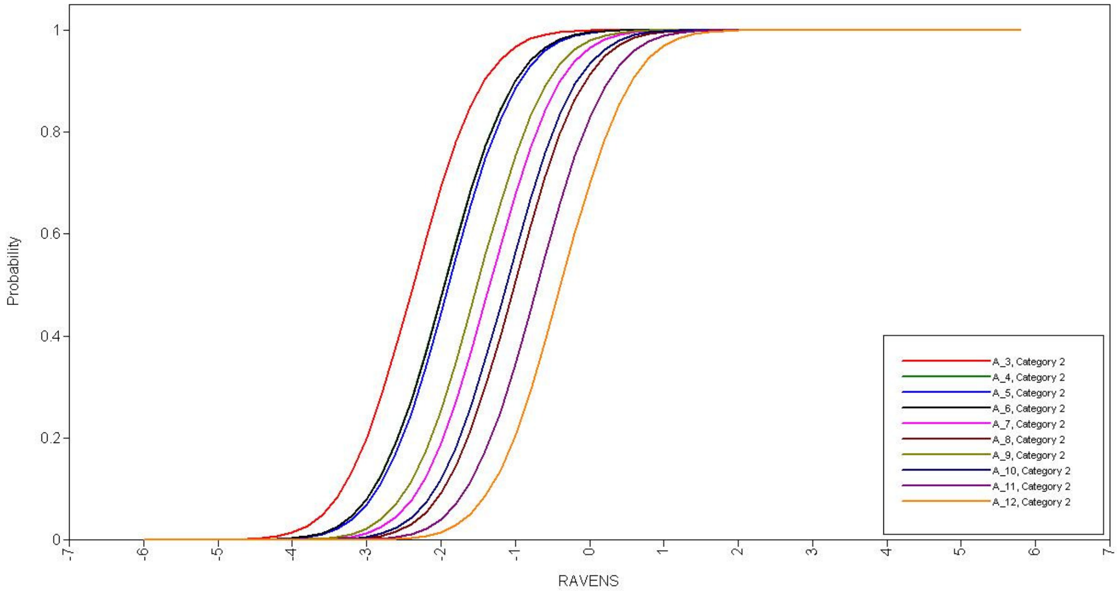


Figure 14: Male Sample Item Characteristic Curves for the Rasch Model of Item Set A of the Raven's SPM+



It is apparent from the results that set A items do not follow a perfect progression of item difficulty. For both males and females, items increase in the following order of difficulty: A3, A4, A6, A5, A9, A7, A10, A8, A11, and A12. However, in terms of average item difficulty (see Table 16), set A items are easier than set B items which will now be discussed further.

5.6.2. Set B

Item difficulty values for items B1 to B12 of set B of the SPM+ are provided in Table 11. These item difficulties are graphically presented in Figure 15 for females and in Figure 16 for males.

Table 11: Item difficulties of Items in Set B for Males and Females

	Male				**Female**			
SPM+ Items	**Est.**	**S.E.**	**Est./S.E.**	**2-tailed p-value**	**Est.**	**S.E.**	**Est./S.E.**	**2-tailed p-value**
B1	-2.564	0.184	-13.896	0.00	-2.338	0.203	-11.543	0.00
B2	-2.591	0.191	-13.571	0.00	-2.359	0.196	-12.02	0.00
B3	-2.466	0.190	-12.974	0.00	-2.264	0.189	-11.955	0.00
B4	-1.500	0.118	-12.757	0.00	-1.529	0.100	-15.215	0.00
B5	-1.743	0.145	-12.054	0.00	-1.714	0.105	-16.382	0.00
B6	-1.006	0.094	-10.660	0.00	-1.153	0.091	-12.72	0.00
B7	-0.905	0.091	-9.9170	0.00	-1.076	0.096	-11.23	0.00
B8	-1.115	0.095	-11.700	0.00	-1.236	0.092	-13.422	0.00
B9	-1.234	0.103	-12.027	0.00	-1.326	0.094	-14.063	0.00
B10	-1.511	0.116	-12.996	0.00	-1.537	0.101	-15.254	0.00
B11	-1.050	0.095	-10.994	0.00	-1.186	0.099	-11.988	0.00
B12	-0.489	0.084	-5.8460	0.00	-0.759	0.122	-6.245	0.00
Average	-1.515				-1.540			

Figure 15: Female sample Item Characteristic Curves for the Rasch Model of Item Set B of the Raven's SPM+

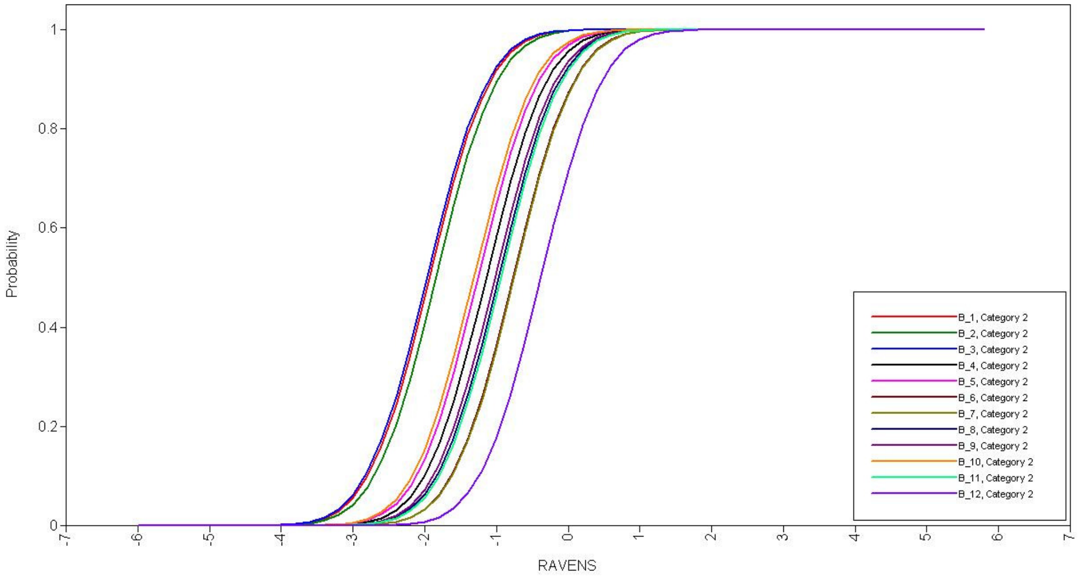
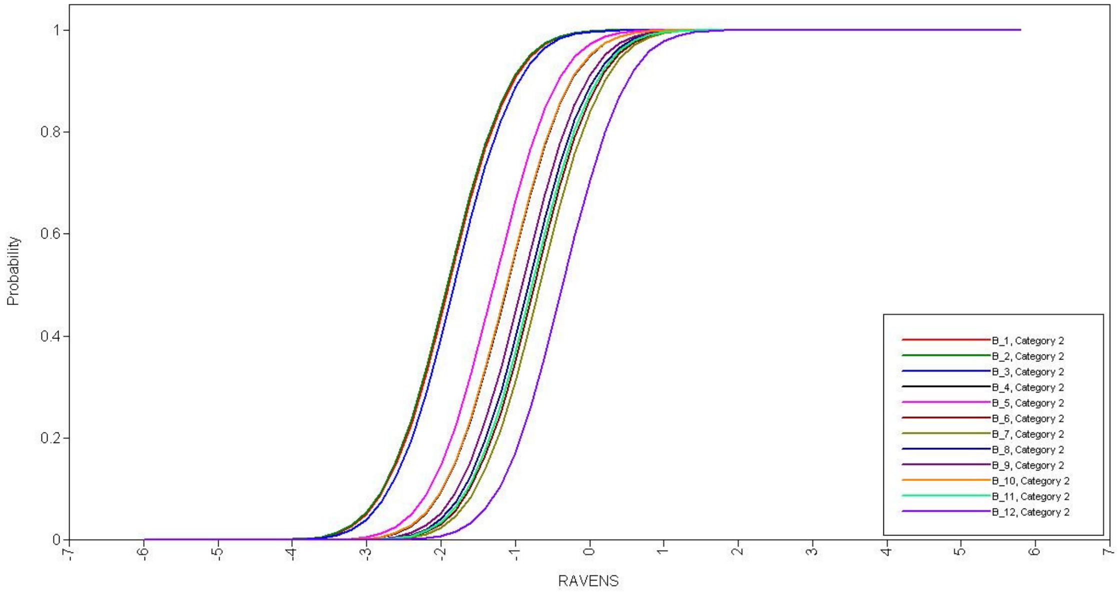


Figure 16: Male sample Item Characteristic Curves for the Rasch Model of Item Set B of the Raven's SPM+



As with item set A, items in set B are not following a perfect progression of item difficulty, from easiest to most difficult. For both males and females, the items, in increasing order of difficulty, are as follows: B2, B1, B3, B5, B10, B4, B9, B8, B11, B6, B7, and B12. This is evidenced by the item difficulty values and ICCs. In terms of average item difficulty (see Table 16), set B items are more difficult than set A, and easier than items in set C which will now be discussed.

5.6.3. Set C

Item difficulty values for items C1 to C12 of set C of the SPM+ are provided in Table 12. These item difficulties are graphically presented in Figure 17 for females and in Figure 18 for males.

Table 12: Item difficulties of Items in Set C for Males and Females

	Male				Female			
SPM+ Items	**Est.**	**S.E.**	**Est./S.E.**	**2-tailed p-value**	**Est.**	**S.E.**	**Est./S.E.**	**2-tailed p-value**
C1	-1.925	0.159	-12.099	0.00	-2.002	0.172	-11.666	0.00
C2	-2.155	0.165	-13.035	0.00	-2.293	0.214	-10.694	0.00
C3	-1.104	0.106	-10.453	0.00	-0.963	0.101	-9.546	0.00
C4	-2.376	0.190	-12.499	0.00	-2.572	0.235	-10.928	0.00
C5	-0.623	0.095	-6.540	0.00	-0.356	0.109	-3.278	0.00
C6	-1.462	0.124	-11.752	0.00	-1.416	0.129	-10.963	0.00
C7	0.441	0.104	4.235	0.00	0.991	0.179	5.524	0.00
C8	-0.446	0.090	-4.974	0.00	-0.131	0.081	-1.614	0.11
C9	1.381	0.133	10.368	0.00	2.179	0.275	7.911	0.00
C10	2.315	0.176	13.146	0.00	3.360	0.383	8.770	0.00
C11	2.423	0.179	13.554	0.00	3.497	0.392	8.921	0.00
C12	1.706	0.148	11.569	0.00	2.591	0.309	8.373	0.00
Average	-0.152				0.240			

Figure 17: Female sample Item Characteristic Curves for the Rasch Model of Item Set C of the Raven's SPM+

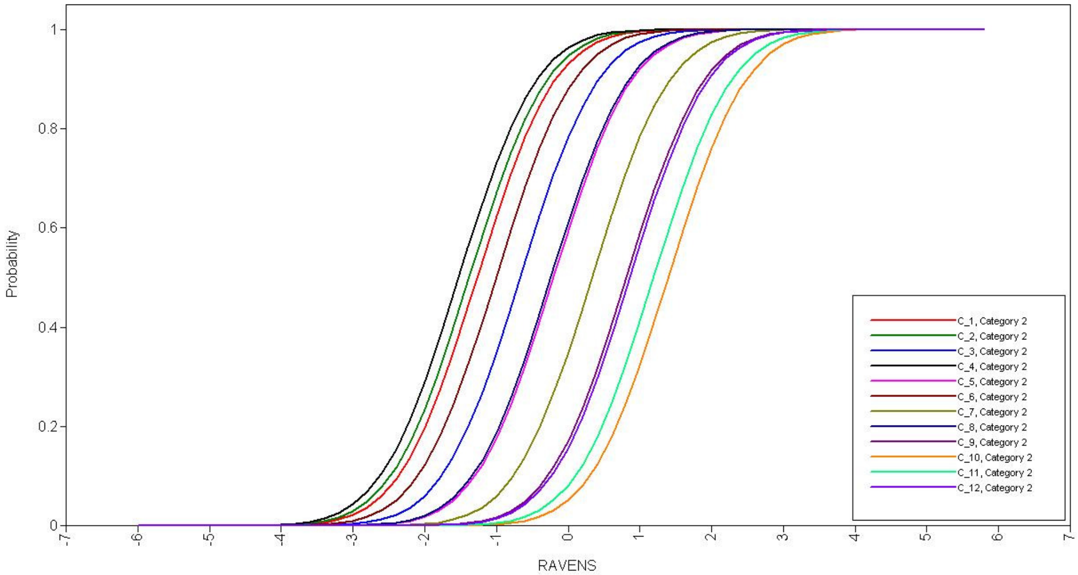
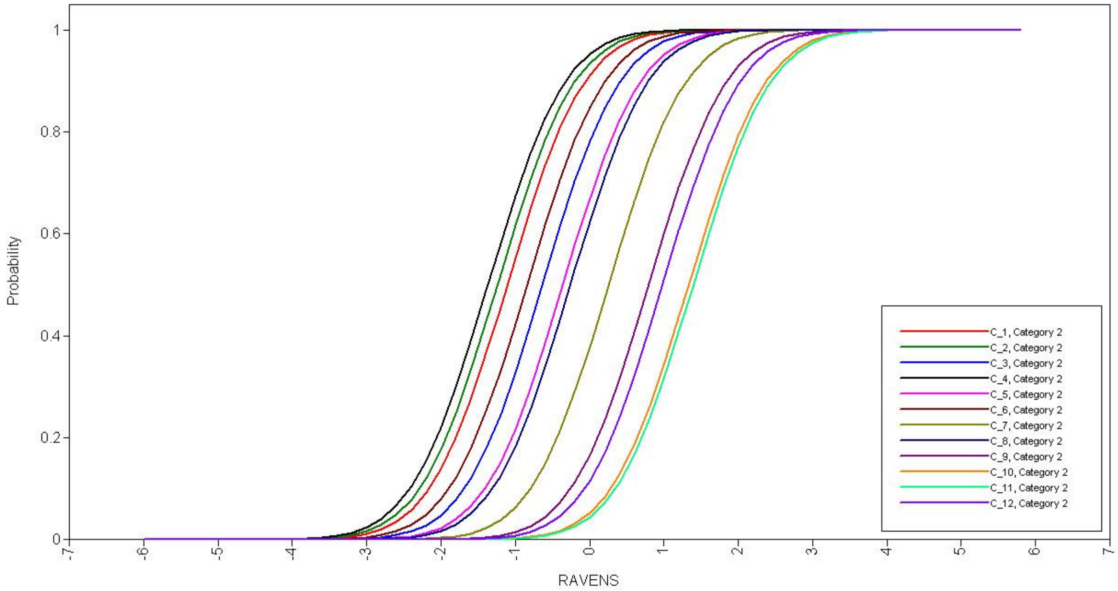


Figure 18: Male sample Item Characteristic Curves for the Rasch Model of Item Set C of the Raven's SPM+



As with item sets A and B, item set C does not begin with the easiest item, nor do the items follow a sequential progression of item difficulty (as evidenced by the item difficulty values and the ICCs). For males and females, the items increase in difficulty in the following order: C4, C2, C1, C6, C3, C5, C8, C7, C9, C12, C10, and C11. The average item difficulty of set C is more difficult than set A and B, but easier than sets D and E.

5.6.4. Set D

Item difficulty values for items D1 to D12 of set D of the SPM+ are provided in Table 13. These item difficulties are graphically presented in Figure 19 for females and in Figure 20 for males.

Table 13: Item difficulties of Items in Set D for Males and Females

	Male				Female			
SPM+ Items	Est.	S.E.	Est./S.E.	2-tailed p-value	Est.	S.E.	Est./S.E.	2-tailed p-value
D1	-0.432	0.107	-4.050	0.000	-0.605	0.104	-5.794	0.000
D2	-0.021	0.093	-0.223	0.824	-0.052	0.068	-0.765	0.444
D3	0.150	0.093	1.607	0.108	0.178	0.087	2.053	0.040
D4	1.440	0.154	9.344	0.000	1.915	0.221	8.661	0.000
D5	1.853	0.183	10.121	0.000	2.469	0.248	9.945	0.000
D6	1.644	0.171	9.611	0.000	2.188	0.255	8.572	0.000
D7	1.094	0.136	8.067	0.000	1.449	0.167	8.670	0.000
D8	2.210	0.203	10.906	0.000	2.950	0.295	9.986	0.000
D9	1.580	0.163	9.684	0.000	2.103	0.247	8.519	0.000
D10	2.084	0.197	10.581	0.000	2.781	0.295	9.439	0.000
D11	2.073	0.201	10.305	0.000	2.766	0.297	9.320	0.000
D12	2.788	0.230	12.110	0.000	3.728	0.382	9.765	0.000
Average	1.372				1.823			

Figure 19: Female sample Item Characteristic Curves for the Rasch Model of Item Set D of the Raven's SPM+

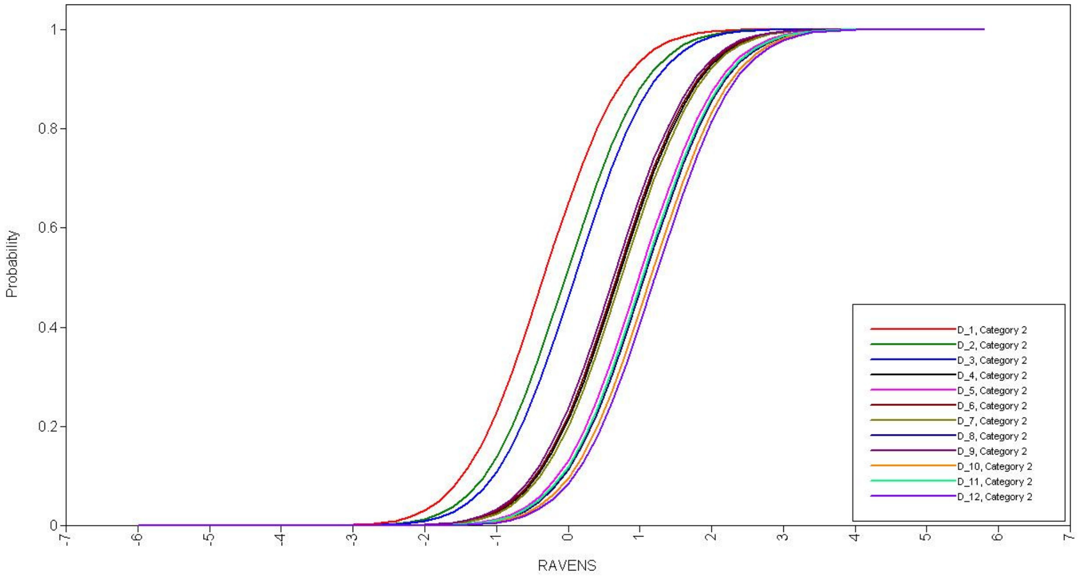
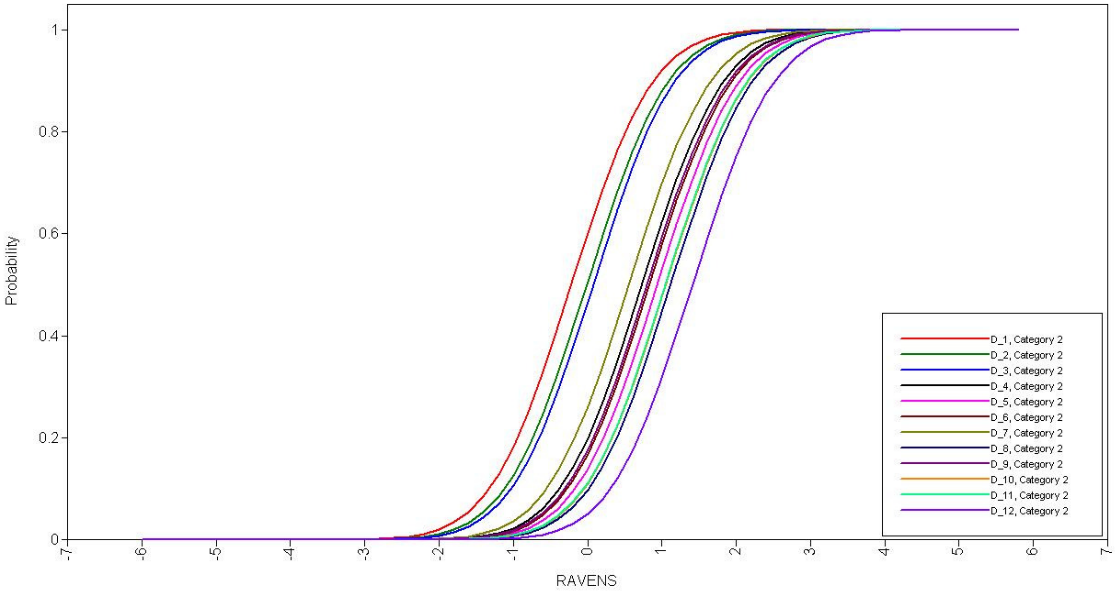


Figure 20: Male sample Item Characteristic Curves for the Rasch Model of Item Set D of the Raven's SPM+



As with previous item sets discussed, items in set D do not follow a perfect progression of item difficulty. The items increase in difficulty, for both males and females, in the following order: D1, D2, D3, D7, D4, D9, D6, D5, D11, D10, D8, and D12. However, the average item difficulty of set D is more difficult than set A, B, and C, but easier than set E.

5.6.5. Set E

Item difficulty values for items E1 to E12 of set E of the SPM+ are provided in Table 14. These item difficulties are graphically presented in Figure 21 for females and in Figure 22 for males.

Table 14: Item difficulties of Items in Set E for Males and Females

SPM+ Items	Male				Female			
	Est.	S.E.	Est./S.E.	2-tailed p-value	Est.	S.E.	Est./S.E.	2-tailed p-value
E1	1.558	0.201	7.758	0.000	0.792	0.169	4.694	0.000
E2	1.798	0.209	8.581	0.000	2.001	0.234	8.554	0.000
E3	1.756	0.208	8.432	0.000	1.462	0.197	7.415	0.000
E4	1.833	0.213	8.607	0.000	2.518	0.277	9.093	0.000
E5	4.026	0.452	8.899	0.000	4.227	0.404	10.450	0.000
E6	3.189	0.350	9.102	0.000	2.958	0.292	10.130	0.000
E7	3.650	0.404	9.042	0.000	3.009	0.307	9.791	0.000
E8	4.159	0.424	9.818	0.000	4.097	0.402	10.196	0.000
E9	4.441	0.435	10.217	0.000	3.648	0.357	10.220	0.000
E10	3.787	0.408	9.287	0.000	3.936	0.381	10.329	0.000
E11	3.958	0.428	9.238	0.000	4.139	0.401	10.321	0.000
E12	4.604	0.463	9.935	0.000	4.097	0.392	10.462	0.000
Average	3.230				3.074			

Figure 21: Female sample Item Characteristic Curves for the Rasch Model of Item Set E of the Raven's SPM+

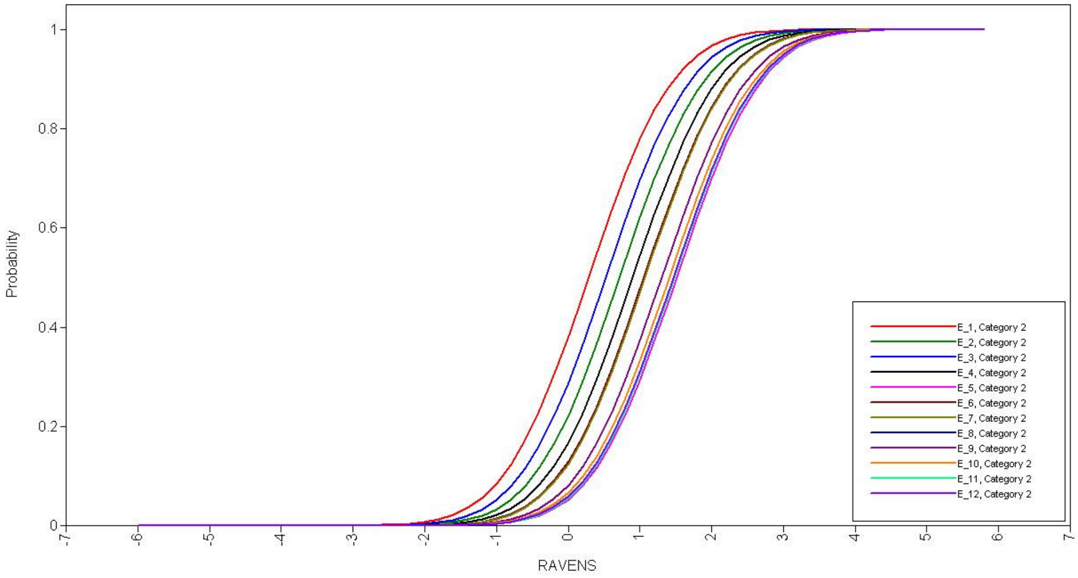
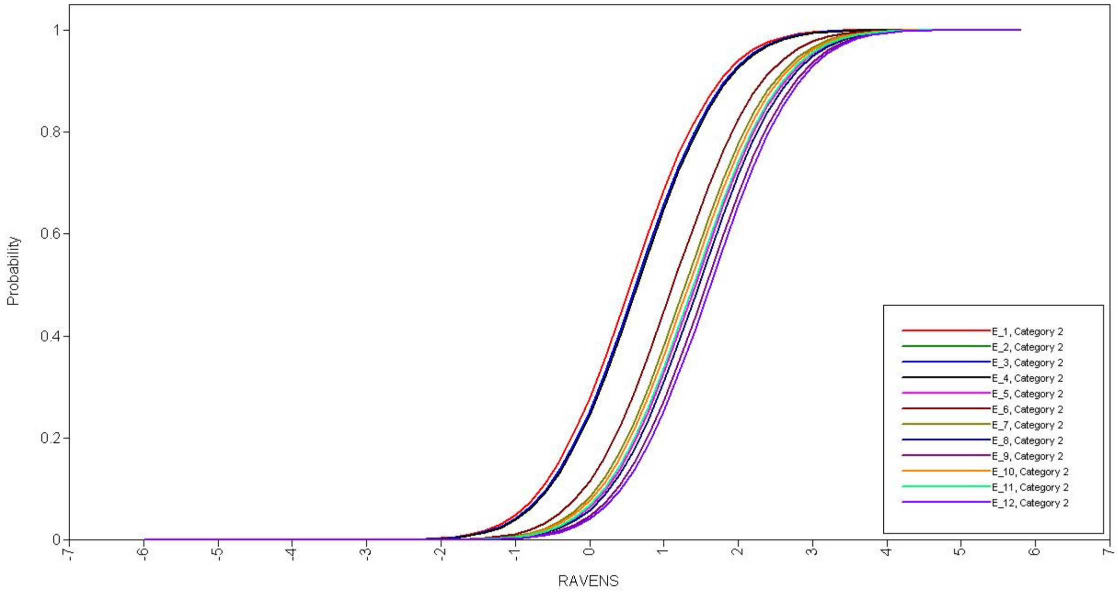


Figure 22: Male sample Item Characteristic Curves for the Rasch Model of Item Set E of the Raven's SPM+



As with the previous item sets, set E items do not follow a perfect progression of item difficulty from beginning to the end of the set. Unlike the previous item sets, the order of item difficulty differs for males and females. For males, the items increase in the following order of difficulty: E1, E3, E2, E4, E6, E7, E5, E11, E10, E8, E9, and E12. For females, the order of items in terms of difficulty is: E1, E3, E2, E4, E6, E7, E9, E10, E8, E12, E11, and

E5. Compared with the other item sets, set E is the most difficult of all the item sets for both males and females as evidenced from the average item difficulty, which fulfils the original design goals of the measure.

Item difficulties from all item sets are presented in Table 15 in sequential order of difficulty, the easiest to most difficult for both males and females. It can be noted that item difficulties do not follow the order in which they are presented, and the sequence of item difficulties is slightly different for males and females

Table 15: Items of the SPM+ for males and females ordered from easiest to most difficult.

	Overall Sample		Males		Females
SPM+ Item	**Item Difficulty**	**SPM+ Item**	**Item Difficulty**	**SPM+ Item**	**Item Difficulty**
A3	-4.336	**A3**	-3.539	**A3**	-3.622
A4	-4.229	**A4**	-2.919	**A4**	-2.978
A6	-4.089	**A6**	-2.911	**A6**	-2.971
A5	-3.891	**A5**	-2.828	**A5**	-2.884
B1	-3.757	**B2**	-2.591	**C4**	-2.572
B2	-3.757	**B1**	-2.564	**B2**	-2.359
B3	-3.726	**B3**	-2.466	**B1**	-2.338
C4	-2.999	**C4**	-2.376	**C2**	-2.293
A9	-2.919	**A9**	-2.250	**A9**	-2.285
B5	-2.688	**C2**	-2.155	**B3**	-2.264
C2	-2.648	**A7**	-2.005	**A7**	-2.03
A7	-2.635	**C1**	-1.925	**C1**	-2.002
C1	-2.572	**B5**	-1.743	**B5**	-1.714
B10	-2.408	**A10**	-1.672	**A10**	-1.685
B4	-2.289	**A8**	-1.511	**B10**	-1.537
A10	-2.198	**B10**	-1.511	**B4**	-1.529
A8	-2.12	**B4**	-1.500	**A8**	-1.518
B9	-1.913	**C6**	-1.462	**C6**	-1.416
C6	-1.862	**B9**	-1.234	**B9**	-1.326

	Overall Sample		Males		Females
SPM+ Item	**Item Difficulty**	**SPM+ Item**	**Item Difficulty**	**SPM+ Item**	**Item Difficulty**
B8	-1.796	**B8**	-1.115	**B8**	-1.236
B6	-1.617	**C3**	-1.104	**B11**	-1.186
B11	-1.61	**B11**	-1.050	**B6**	-1.153
B7	-1.415	**A11**	-1.049	**B7**	-1.076
A11	-1.34	**B6**	-1.006	**A11**	-1.038
C3	-1.34	**B7**	-0.905	**C3**	-0.963
B12	-0.737	**C5**	-0.623	**B12**	-0.759
A12	-0.685	**A12**	-0.579	**D1**	-0.605
D1	-0.599	**B12**	-0.489	**A12**	-0.55
C5	-0.509	**C8**	-0.446	**C5**	-0.356
C8	-0.261	**D1**	-0.432	**C8**	-0.131
D2	-0.06	**D2**	-0.021	**D2**	-0.052
D3	0.201	**D3**	0.150	**D3**	0.178
E1	0.577	**C7**	0.441	**E1**	0.792
C7	0.588	**D7**	1.094	**C7**	0.991
E3	1.086	**C9**	1.381	**D7**	1.449
D7	1.306	**D4**	1.440	**E3**	1.462
D9	1.387	**E1**	1.558	**D4**	1.915
D4	1.478	**D9**	1.580	**E2**	2.001
D6	1.493	**D6**	1.644	**D9**	2.103
C9	1.602	**C12**	1.706	**C9**	2.179
E2	1.602	**E3**	1.756	**D6**	2.188
E4	1.725	**E2**	1.798	**D5**	2.469
C12	1.871	**E4**	1.833	**E4**	2.518
D5	1.965	**D5**	1.853	**C12**	2.591
D8	2.139	**D11**	2.073	**D11**	2.766
E6	2.139	**D10**	2.084	**D10**	2.781
D10	2.149	**D8**	2.210	**D8**	2.95
D11	2.168	**C10**	2.315	**E6**	2.958
E7	2.342	**C11**	2.423	**E7**	3.009

	Overall Sample		Males		Females
SPM+ Item	**Item Difficulty**	**SPM+ Item**	**Item Difficulty**	**SPM+ Item**	**Item Difficulty**
C11	2.453	**D12**	2.788	**C10**	3.36
D12	2.584	**E6**	3.189	**C11**	3.497
C10	2.648	**E7**	3.650	**E9**	3.648
E10	2.77	**E10**	3.787	**D12**	3.728
E9	2.857	**E11**	3.958	**E10**	3.936
E11	2.857	**E5**	4.026	**E8**	4.097
E8	2.903	**E8**	4.159	**E12**	4.097
E5	2.934	**E9**	4.441	**E11**	4.139
E12	3.05	**E12**	4.604	**E5**	4.227

Despite the varying item-level difficulties, if the average item difficulty of each item set is considered (Table 16), each successive item set is progressively more difficult across the SPM+, and follows the same order for both males and females.

Table 16: Average item difficulty of all item sets for males and females

Item Set	Males	Females
A	-2.126	-2.156
B	-1.515	-1.540
C	-0.152	0.240
D	1.372	1.822
E	3.230	3.073

Evidence from the Rasch model indicates that the item sets are increasing in difficulty despite the fact that within the item sets there is not a sequential progression of difficulty of the items themselves.

5.7. Summary

The factorial structure of the SPM+ has long been disputed. Results of the EFA of the U.K. standardisation data suggest that the data can be described sufficiently well by a one-factor model and supports a main research objective of this dissertation. In this chapter, a one-factor model was assessed at the item level through the use of the Rasch model.

From the results of the Rasch model, it can be further concluded that a one-factor model adequately fits the data from the SPM+ as evidenced by the item difficulty values and the increase of average item difficulties across item sets. Raven (2009) acknowledged that the development of a “perfect” progression of item difficulties would not be possible, but that rather, the average would be a reasonable index of ability.

One is reminded that Rasch models are limited in that they provide an account of only one parameter of the model, item difficulty, while assuming item discrimination is constant across all items. With respect to the overarching question of this thesis it is important to consider other parameters of the model that could affect the understanding of sex differences in general intelligence.

Using a two-parameter logistic model, it is possible to model both item difficulty and item discrimination simultaneously. This allows for a parsimonious yet robust assessment of group differences, resulting in the ability to draw more confident conclusions about sex differences in general intelligence. The two-parameter normal ogive item response theory model (i.e., 2PL) is known to be equivalent to confirmatory factor analysis with binary outcomes (Brown, 2006), and aims to explain correlations among a set of test items. In the next chapter, unidimensionality of the Raven’s SPM+ will be assessed using confirmatory factor analysis methods. Upon establishment of a suitable CFA model, sex differences in mean and variability will be assessed.

— 6 —

Results Chapter 2:

Assessment of Group Differences in Mean and Variability Using Multiple Groups Confirmatory Factor Analysis

6.1. Introduction

The first two objectives of this dissertation will be examined in this chapter: Is the SPM+ measuring general intelligence in the same way for males and females?; and Are there sex differences in the overall sample of the SPM+? A further dissertation objective will also be addressed in this chapter: Are the results of the overall sample of the SPM+ being affected by methods effects?

As previously discussed, the existence of sex differences in general intelligence has long been debated by researchers and the general public alike in an attempt to resolve the question 'which is the smarter sex?' Results to date have been inconsistent: males outperform females in some studies (Colom, Escorial, & Rebollo, 2004; Irwing & Lynn, 2005; Jackson & Rushton, 2006; Lynn, 1994; 1998; Mackintosh & Bennett, 2005; Silverman et al., 2000; Vigneau & Bors, 2008), females outperform males (Abdel-Khalek & Lynn, 2006; Khaleefa & Lynn, 2008a) while in other studies, no mean sex differences have been found (Crucian & Berenbaum, 1998; Khaleefa & Lynn, 2008b; Rushton & Cvorovic, 2009). Reasons for the inconsistent findings are unclear, however one potential reason may be that studies have not generally used modern psychometric methods that allow for the appropriate assessment of item bias and that account for measurement error while assessing for mean differences in the latent trait of general intelligence. These features are essential to reliably compare group differences.

As detailed in Chapter 4, classical methods for the assessment of group differences are no longer considered sufficiently sensitive to correctly identify the sources of bias that may occur within data. This could result in a Type I error, or rejecting the null hypothesis when it is true. In addition, classical methods may also falsely identify differences that do not truly exist. This would lead to the commission of a Type II error, or the acceptance of the null hypothesis when it is false. Unlike modern latent variable modelling, classical methods fail to account for measurement error. Further, when conducting group comparisons, classical methods do not allow for the item parameters to be held to equality in order to identify the source of any potential differences (Brown, 2006).

As detailed in Chapter 5, a one-factor model can be used to effectively explain the UK standardisation data of the Raven's SPM+. A Rasch model was employed, where the discrimination and the guessing parameter of each item were held to equality across the item subsets while the item difficulty was allowed to be freely estimated. This allowed for an assessment of the interplay between the latent construct of general intelligence of the respondents of the U.K. standardisation sample and the characteristics of the items on the SPM+. However, it is also important to fully investigate the item discrimination as well as the item difficulty.

Item discrimination is the parameter that determines the likelihood of answering the item correctly in relation to the respondents' level of general intelligence. In Item Response Theory terminology, a two-parameter logistic model (2PL) allows the item difficulty and item discrimination to be freely estimated while the guessing parameter is held to equality. With respect to binary data, as is the case with the SPM+, a 2PL normal ogive IRT model is known to be equivalent to confirmatory factor analysis. Item discrimination parameters are equivalent to factor loadings in CFA, and represent the relationship between the latent trait and the item response (Brown, 2006; Embretson & Reise, 2000; MacDonald, 1999).

One type of confirmatory factor analysis particularly suited to the assessment of group differences is multi-group confirmatory factor analysis (MG-CFA). MG-CFA allows for the simultaneous confirmatory factor analysis of more than one group. For example, employing MG-CFA techniques allows for the assessment of item difficulty and discrimination for males and females separately in the same analysis. In addition, MG-CFA allows group comparisons of measurement characteristics (i.e., factor loadings, indicator intercepts or thresholds, residual variances) and the structural features (i.e., factor variances, factor covariances, and latent means) of the latent factor(s) which is of utmost importance to ensure that the measurement properties are equivalent in each group (Brown, 2006; van Der Sluis et al., 2008). In contrast, other forms of latent variable modelling (such as Multiple Indicator Multiple Causes or MIMIC) can only assess group differences in indicator thresholds and latent means.

Measurement invariance exists when the relation between the observed test scores and the underlying latent attribute are the same in both groups (van der Sluis et al., 2008; Drasgow, 1984; Horn & McArdle, 1992). It ensures that any claims that are made about group differences, or lack thereof, are genuine and not attributable to artefacts arising from measurement error or bias.

When models are nested, or aspects of the model specification are layered upon on another (as with the assessment of measurement invariance), the chi-square statistic can be used to statistically compare the model fit. It is imperative that the models satisfy the requirements of adequate model fit in their own rights before being compared. When the difference between the chi-squares of two models is assessed, it is referred to as the *chi-square difference test* (Brown, 2006). Once the difference of chi-square between the models has been tabulated, the resulting value is compared to a critical value. If the chi-square difference value exceeds the critical values, then it can be concluded that the second model presents significantly better fit.

Upon evaluation of the models, it is important to consider alternative explanations for residual variance and model fit. When some of the differential covariance amongst a set of items is attributable to the measurement approach rather than the latent construct, it is often referred to as a method effect (Brown, 2006).

As discussed in Chapter 4, methods effects can arise from the modality of assessment (such as items presented in questionnaire or multiple choice format) or may be due to the way items are worded or presented (Brown, 2006). The covariation reflects an artefact of different response styles associated with the way in which the item is presented, and is not based upon different dimensions of the underlying latent factor.

For example, Marsh (1996) challenged the commonly-used two-factor solution of the Self-Esteem Questionnaire (SEQ) through the use of a single factor solution with a method effect. Traditionally, the positively- and negatively-worded items of the questionnaire were conceptualised as two distinct factors. However, Marsh determined that the differential covariance among the items was not based upon substantively different dimensions, but rather the covariation was due to the directionality of the item wording.

Practically speaking, when an indicator is specified to load onto 2 factors simultaneously (such as a general intelligence factor and a methods factor, as will be seen in this chapter), the indicator is loaded onto the first factor, with the residual variance of that indicator loaded onto the second factor. Similar to the positive- and negatively-worded items of the SEQ, it is possible that elements of the SPM+ items (such as the solving strategy required to solve the item or figural elements of the item) may be accounting for some of the covariation. In the conceptualisation of the methods effects analyses, the existing literature pertaining to the multidimensionality of the Raven's (Carpenter, Just, & Shell, 1990; Deshon, Chan, & Weissbein, 1995; Lynn, Allik, & Irwing, 2004; van der Ven & Ellis, 2000) will be referenced.

The overall objective of this chapter is to compare boys' and girls' performance on the SPM+, and will be done using two approaches. First, a one-factor Confirmatory Factor Analysis model of general intelligence is established for males and females separately (section 6.2.1). This will further verify previous claims about the suitability of a one-factor model of general intelligence for males and females separately. Additionally, it allows for detailed item-level assessment of difficulty and discrimination for males and females. Second, a MG-CFA model is established where two separate correlation input matrices will be analysed simultaneously for boys and girls (section 6.2.12). Using this model, the equivalence of the measurement properties of the SPM+ latent factor for girls and boys are examined. The latent means of the SPM+ latent factor for boys and girls are then compared, and the variability of the SPM+ latent factor for boys and girls is evaluated (section 6.3). Finally, in an attempt to further account for unexplained variance in the models, methods factors will be tested (section 6.4).

6.2. Model Specification & Measurement Invariance

To determine how well the items on the Raven's SPM+ can be explained by one underlying latent trait of general intelligence using a two parameter logistic model, a unidimensional confirmatory factor analysis was initially conducted (Brown, 2006). The 58 SPM+ test items were loaded onto one latent factor and analysed in two separate input correlation matrices, one for males (Table 17) and one for females (Table 18). A diagram of the one-factor model that was used for both males and females is provided below (Figure 23). Due to the number of indicators in the model, it was not possible to include all of the relevant information in the diagram. Therefore, standardised and unstandardised factor loadings and error variances are available in Table 19.

Table 17. Correlation Matrix for Male 1-factor model

<Insert Table *** April 6 2010 Correlation Matrices 1-factor model BOY GIRL.xlsx>

Table 18. Correlation Matrix for Female 1-factor model

<Insert Table *** April 6 2010 Correlation Matrices 1-factor model BOY GIRL.xlsx>

Figure 23: Diagram of the One-Factor MG-CFA for Males and Females

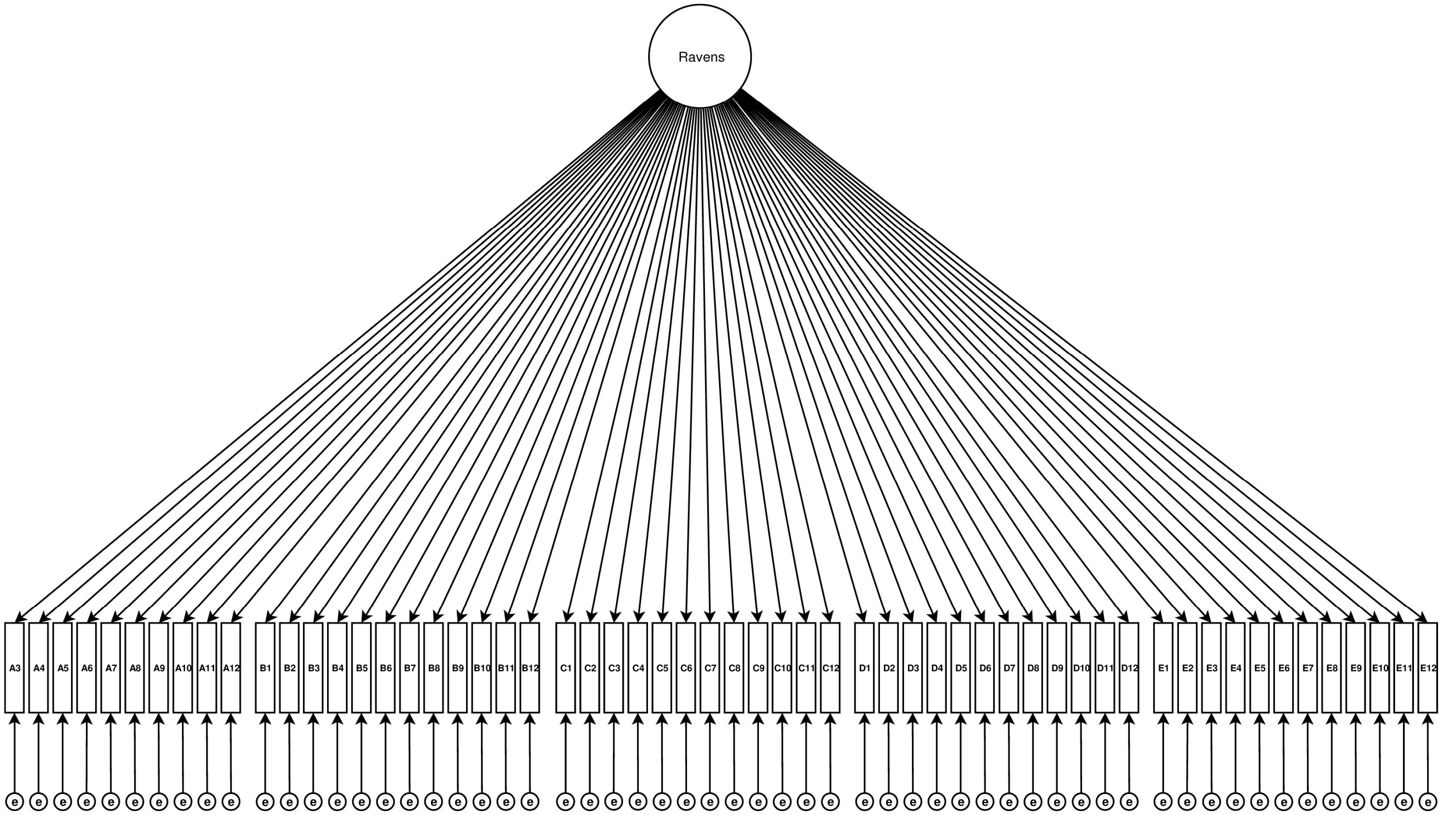


Table 19: Unstandardised Parameter Estimates, Indicator Thresholds, and Proportion of Variance Explained (R^2) for Males and Females

	Male		Female		Male		Female		Male		Female		Male		Female		Male		Female	
Indicator	Unstand. Factor Loading	*p*	Unstand. Factor Loading	*p*	Standardised Factor Loading	*p*	Standardised Factor Loading	*p*	Unstand. Indicator Threshold	*p*	Unstand. Indicator Threshold	*p*	Standardised Indicator Threshold	*p*	Standardised Indicator Threshold	*p*	Prop. of Variance Explained (R^2)	*p*	Prop. of Variance Explained (R^2)	*p*
A3	1.000	0.000	1.000	0.000	0.531	0.000	0.592	0.000	-1.993	0.000	-1.993	0.000	-1.993	0.000	-2.267	0.000	0.282	0.000	0.351	0.000
A4	1.049	0.000	1.049	0.000	0.557	0.000	0.542	0.000	-2.102	0.000	-2.102	0.000	-2.102	0.000	-2.088	0.000	0.311	0.000	0.294	0.010
A5	1.058	0.000	1.058	0.000	0.562	0.000	0.648	0.000	-1.765	0.000	-1.765	0.000	-1.765	0.000	-2.078	0.000	0.316	0.000	0.420	0.000
A6	1.161	0.000	1.161	0.000	0.617	0.000	0.723	0.000	-1.842	0.000	-1.842	0.000	-1.842	0.000	-2.206	0.000	0.380	0.000	0.523	0.000
A7	1.030	0.000	1.030	0.000	0.547	0.000	0.502	0.000	-1.350	0.000	-1.350	0.000	-1.350	0.000	-1.264	0.000	0.299	0.000	0.252	0.000
A8	1.281	0.000	1.281	0.000	0.680	0.000	0.613	0.000	-1.099	0.000	-1.099	0.000	-1.099	0.000	-1.010	0.000	0.463	0.000	0.375	0.000
A9	1.345	0.000	1.345	0.000	0.714	0.000	0.512	0.000	-1.650	0.000	-1.650	0.000	-1.650	0.000	-1.207	0.000	0.510	0.000	0.262	0.000
A10	1.343	0.000	1.343	0.000	0.714	0.000	0.610	0.000	-1.175	0.000	-1.175	0.000	-1.175	0.000	-1.026	0.000	0.509	0.000	0.372	0.000
A11	1.069	0.000	1.069	0.000	0.568	0.000	0.514	0.000	-0.691	0.000	-0.691	0.000	-0.691	0.000	-0.639	0.000	0.322	0.000	0.264	0.000
A12	0.704	0.000	0.704	0.000	0.374	0.000	0.443	0.000	-0.301	0.000	-0.301	0.000	-0.301	0.000	-0.363	0.000	0.140	0.000	0.196	0.000
B1	1.111	0.000	1.111	0.000	0.590	0.000	0.580	0.000	-1.865	0.000	-1.865	0.000	-1.865	0.000	-1.872	0.000	0.348	0.000	0.337	0.000
B2	1.023	0.000	1.023	0.000	0.543	0.000	0.691	0.000	-1.570	0.000	-1.570	0.000	-1.570	0.000	-2.039	0.000	0.295	0.000	0.478	0.000
B3	1.568	0.000	1.300	0.000	0.833	0.000	0.748	0.000	-1.764	0.000	-1.764	0.000	-1.764	0.000	-1.950	0.000	0.694	0.000	0.559	0.000
B4	1.412	0.000	1.412	0.000	0.750	0.000	0.781	0.000	-1.096	0.000	-1.096	0.000	-1.096	0.000	-1.165	0.000	0.563	0.000	0.610	0.000
B5	1.679	0.000	1.679	0.000	0.892	0.000	0.824	0.000	-1.375	0.000	-1.375	0.000	-1.375	0.000	-1.296	0.000	0.796	0.000	0.678	0.000
B6	1.322	0.000	1.322	0.000	0.702	0.000	0.720	0.000	-0.777	0.000	-0.777	0.000	-0.777	0.000	-0.813	0.000	0.493	0.000	0.518	0.000
B7	0.985	0.000	0.985	0.000	0.523	0.000	0.635	0.000	-0.619	0.000	-0.619	0.000	-0.619	0.000	-0.767	0.000	0.274	0.000	0.403	0.000
B8	1.352	0.000	1.352	0.000	0.718	0.000	0.760	0.000	-0.854	0.000	-0.854	0.000	-0.854	0.000	-0.922	0.000	0.516	0.000	0.577	0.000
B9	1.239	0.000	1.239	0.000	0.658	0.000	0.662	0.000	-0.935	0.000	-0.935	0.000	-0.935	0.000	-0.960	0.000	0.433	0.000	0.438	0.000
B10	1.512	0.000	1.512	0.000	0.803	0.000	0.819	0.000	-1.168	0.000	-1.168	0.000	-1.168	0.000	-1.216	0.000	0.646	0.000	0.671	0.000
B11	1.053	0.000	1.053	0.000	0.560	0.000	0.556	0.000	-0.789	0.000	-0.789	0.000	-0.789	0.000	-0.801	0.000	0.313	0.000	0.309	0.000
B12	1.003	0.000	1.003	0.000	0.533	0.000	0.507	0.000	-0.367	0.000	-0.367	0.000	-0.367	0.000	-0.356	0.000	0.284	0.000	0.257	0.000
C1	1.503	0.000	1.503	0.000	0.798	0.000	0.737	0.000	-1.316	0.000	-1.316	0.000	-1.316	0.000	-1.240	0.000	0.638	0.000	0.543	0.000
C2	1.064	0.000	1.064	0.000	0.565	0.000	0.592	0.000	-1.271	0.000	-1.271	0.000	-1.271	0.000	-1.360	0.000	0.320	0.000	0.351	0.000
C3	1.145	0.000	1.145	0.000	0.608	0.000	0.643	0.000	-0.633	0.000	-0.633	0.000	-0.633	0.000	-0.684	0.000	0.370	0.000	0.414	0.000
C4	1.403	0.000	1.403	0.000	0.745	0.000	0.745	0.000	-1.473	0.000	-1.473	0.000	-1.473	0.000	-1.504	0.000	0.556	0.000	0.555	0.000
C5	1.203	0.000	1.203	0.000	0.639	0.000	0.513	0.000	-0.278	0.000	-0.278	0.000	-0.278	0.000	-0.228	0.000	0.409	0.000	0.263	0.000

*Item factor loading did not reach the minimum required threshold of 0.3 (McDonald, 1999).
**Item factor loading did not reach statistical significance.

	Male		Female		Male		Female		Male		Female		Male		Female		Male		Female	
Indicator	Unstand. Factor Loading	*p*	Unstand. Factor Loading	*p*	Standardised Factor Loading	*p*	Standardised Factor Loading	*p*	Unstand. Indicator Threshold	*p*	Unstand. Indicator Threshold	*p*	Standardised Indicator Threshold	*p*	Standardised Indicator Threshold	*p*	Prop. of Variance Explained (R^2)	*p*	Prop. of Variance Explained (R^2)	*p*
C6	1.231	0.000	1.231	0.000	0.654	0.000	0.712	0.000	-0.872	0.000	-0.872	0.000	-0.872	0.000	-0.970	0.000	0.427	0.000	0.507	0.000
C7	0.824	0.000	0.824	0.000	0.438	0.000	0.350	0.000	0.322	0.000	0.322	0.000	0.322	0.000	0.263	0.000	0.191	0.000	0.123	0.000
C8	1.094	0.000	1.094	0.000	0.581	0.000	0.621	0.000	-0.121	0.020	-0.121	0.020	-0.121	0.020	-0.132	0.019	0.338	0.000	0.386	0.000
C9	0.831	0.000	0.831	0.000	0.441	0.000	0.400	0.000	0.837	0.000	0.837	0.000	0.837	0.000	0.774	0.000	0.195	0.000	0.160	0.000
C10	0.255	0.030	0.255	0.030	0.136	0.030	0.155*	0.020	1.219	0.000	1.219	0.000	1.219	0.000	1.426	0.000	0.018	0.270	0.024	0.260
C11	0.673	0.000	-0.106*	0.58*	0.358	0.000	-0.052*	0.59**	1.265	0.000	1.265	0.000	1.265	0.000	1.191	0.000	0.128	0.020	0.003	0.790
C12	0.791	0.000	0.791	0.000	0.420	0.000	0.399	0.000	0.957	0.000	0.957	0.000	0.957	0.000	0.926	0.000	0.177	0.000	0.159	0.000
D1	0.966	0.000	0.966	0.000	0.513	0.000	0.617	0.000	-0.261	0.000	-0.261	0.000	-0.261	0.000	-0.321	0.000	0.263	0.000	0.381	0.000
D2	1.001	0.000	1.001	0.000	0.532	0.000	0.709	0.000	-0.017	0.720	-0.017	0.720	-0.017	0.720	-0.023	0.717	0.283	0.000	0.502	0.000
D3	1.081	0.000	1.081	0.000	0.575	0.000	0.583	0.000	0.106	0.040	0.106	0.040	0.106	0.040	0.110	0.039	0.330	0.000	0.340	0.000
D4	0.824	0.000	0.824	0.000	0.438	0.000	0.410	0.000	0.763	0.000	0.763	0.000	0.763	0.000	0.729	0.000	0.192	0.000	0.168	0.000
D5	0.753	0.000	0.753	0.000	0.400	0.000	0.376	0.000	1.008	0.000	1.008	0.000	1.008	0.000	0.967	0.000	0.160	0.000	0.141	0.000
D6	0.587	0.000	0.587	0.000	0.312	0.000	0.297	0.000	0.762	0.000	0.762	0.000	0.762	0.000	0.740	0.000	0.097	0.010	0.088	0.000
D7	0.908	0.000	0.908	0.000	0.483	0.000	0.523	0.000	0.627	0.000	0.627	0.000	0.627	0.000	0.694	0.000	0.233	0.000	0.273	0.000
D8	0.854	0.000	0.854	0.000	0.454	0.000	0.348	0.000	1.194	0.000	1.194	0.000	1.194	0.000	0.936	0.000	0.206	0.000	0.121	0.000
D9	0.514	0.000	0.514	0.000	0.273	0.000	0.240*	0.000	0.739	0.000	0.739	0.000	0.739	0.000	0.663	0.000	0.075	0.010	0.058	0.010
D10	0.807	0.000	0.807	0.000	0.429	0.000	0.325	0.000	1.181	0.000	1.181	0.000	1.181	0.000	0.913	0.000	0.184	0.000	0.106	0.000
D11	0.633	0.000	0.633	0.000	0.336	0.000	0.435	0.000	0.904	0.000	0.904	0.000	0.904	0.000	1.194	0.000	0.113	0.000	0.189	0.000
D12	0.426	0.000	0.426	0.000	0.227*	0.000	0.221*	0.000	1.290	0.000	1.290	0.000	1.290	0.000	1.283	0.000	0.051	0.010	0.049	0.010
E1	0.962	0.000	0.962	0.000	0.511	0.000	0.528	0.000	0.287	0.000	0.287	0.000	0.287	0.000	0.302	0.000	0.261	0.000	0.279	0.000
E2	1.033	0.000	1.033	0.000	0.549	0.000	0.495	0.000	0.845	0.000	0.845	0.000	0.845	0.000	0.778	0.000	0.301	0.000	0.245	0.000
E3	0.867	0.000	0.867	0.000	0.461	0.000	0.471	0.000	0.536	0.000	0.536	0.000	0.536	0.000	0.559	0.000	0.212	0.000	0.222	0.000
E4	1.103	0.000	1.103	0.000	0.586	0.000	0.584	0.000	0.865	0.000	0.865	0.000	0.865	0.000	0.881	0.000	0.343	0.000	0.341	0.000
E5	0.101	0.360	0.101	0.360	0.053*	0.360	0.054*	0.36**	1.444	0.000	1.444	0.000	1.444	0.000	1.489	0.000	0.003	0.650	0.003	0.650
E6	0.003	0.980	0.003	0.980	0.002*	0.98**	0.002*	0.98**	1.083	0.000	1.083	0.000	1.083	0.000	1.057	0.000	0.000	0.990	0.000	0.990
E7	0.203	0.080	0.203	0.080	0.108*	0.07**	0.09*	0.070	1.280	0.000	1.280	0.000	1.280	0.000	1.087	0.000	0.012	0.370	0.008	0.370
E8	0.027	0.790	0.027	0.790	0.014*	0.78**	0.014	0.780	1.432	0.000	1.432	0.000	1.432	0.000	1.468	0.000	0.000	0.890	0.000	0.890
E9	0.218	0.060	0.218	0.060	0.116*	0.06**	0.103*	0.050	1.510	0.000	1.510	0.000	1.510	0.000	1.369	0.000	0.013	0.340	0.011	0.330
E10	0.281	0.020	0.281	0.020	0.149*	0.010	0.132*	0.020	1.454	0.000	1.454	0.000	1.454	0.000	1.312	0.000	0.022	0.210	0.017	0.240
E11	0.259	0.010	0.259	0.010	0.138*	0.000	0.151*	0.000	1.346	0.000	1.346	0.000	1.346	0.000	1.508	0.000	0.019	0.150	0.023	0.130
E12	-0.043	0.710	-0.043	0.710	-0.023*	0.71**	-0.021*	0.71**	1.574	0.000	1.574	0.000	1.574	0.000	1.467	0.000	0.001	0.850	0.000	0.850

*Item factor loading did not reach the minimum required threshold of 0.3 (McDonald, 1999).

**Item factor loading did not reach statistical significance.

For both the males and the females, one-factor models provided good representation of the data. The male one-factor model was over-identified with 172 *df*; χ^2 (172) = 306.371, $p < 0.001$. The statistics indicated good model fit: CFI=0.882, TLI=0.913, RMSEA=0.042. For females, the one-factor model was over-identified with 203 *df*; χ^2 (203) = 337.655, $p < 0.001$. The statistics indicated good model fit: CFI=0.888, TLI=0.929, RMSEA=0.037. Details of the minimum, maximum, and average factor loadings, indicator thresholds and proportion of explained variance (R^2) are available in Table 20.

Table 20: Minimum, Maximum and Average Factor Loadings, Indicator Thresholds, Proportion of Variance Explained (R^2) and Factor Variance for Male and Female 1-factor models

		Male	Female
Factor Loadings	**Min**	0.500	0.296
	Max	1.771	1.309
	Average	1.099	0.849
Indicator Thresholds	**Min**	-2.205	-2.249
	Max	1.599	1.479
	Average	-0.320	-0.360
R^2	**Min**	0.062	0.035
	Max	0.780	0.691
	Average	0.319	0.321
Factor Variance	**Unstand.**	0.282	0.271
	Standard.	1.000	1.000

In addition to the fit statistics, model fit is assessed at the item level by first inspecting the Item Characteristic Curves (ICCs), and secondly, the factor loadings. Item Characteristic Curves for males and females of item sets A through E are available in Figure 24 through Figure 33.

Figure 24: Male Item Characteristic Curves for the MG-CFA of Item Set A

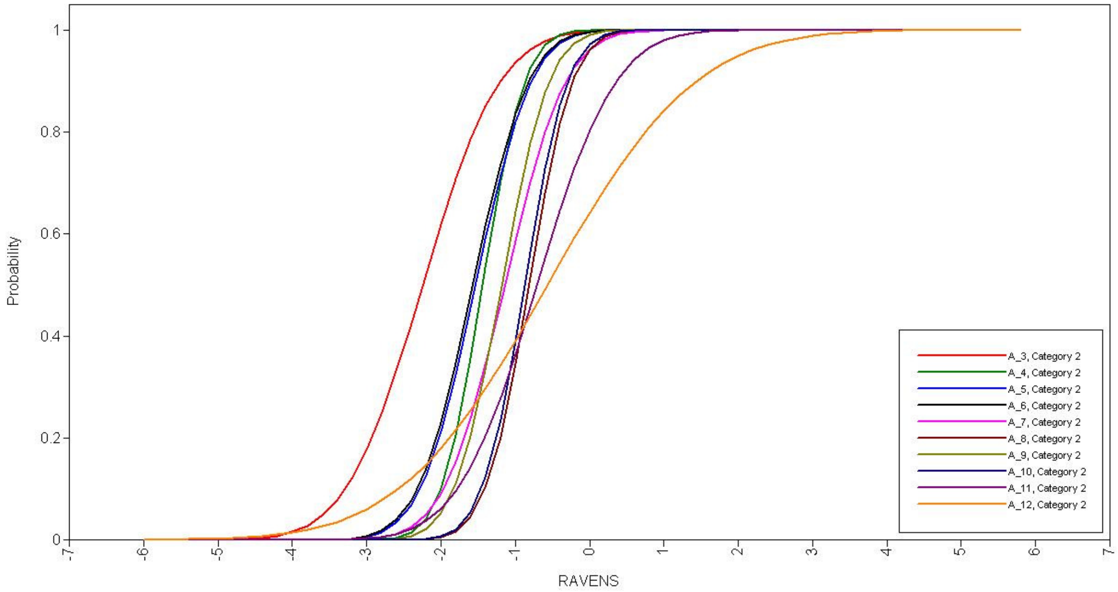


Figure 25: Female Item Characteristic Curves for the MG-CFA of Item Set A

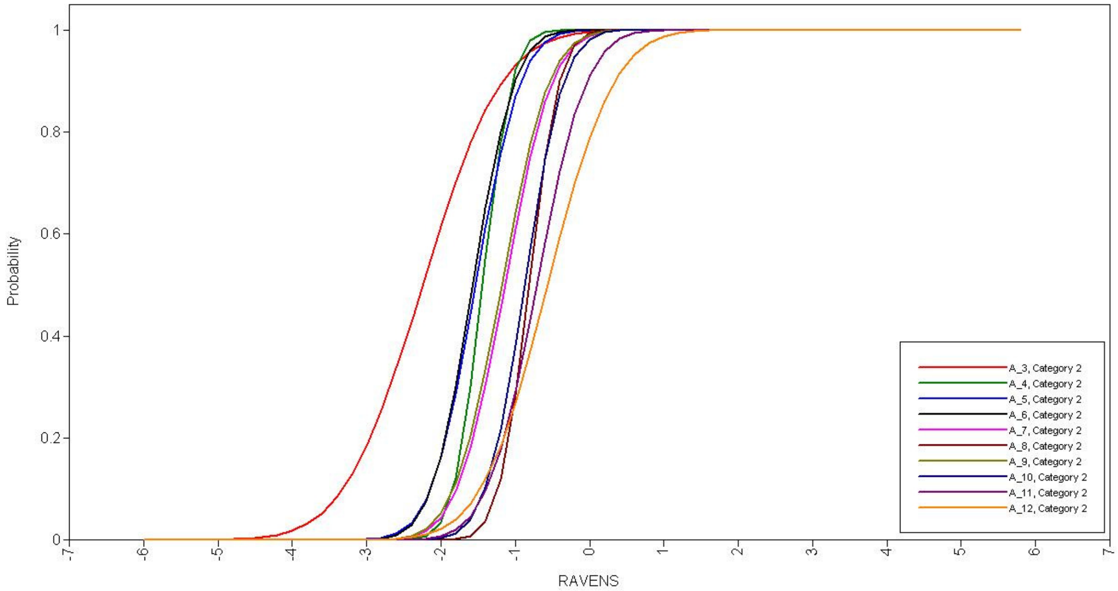


Figure 26: Male Item Characteristic Curves for the MG-CFA of Item Set B

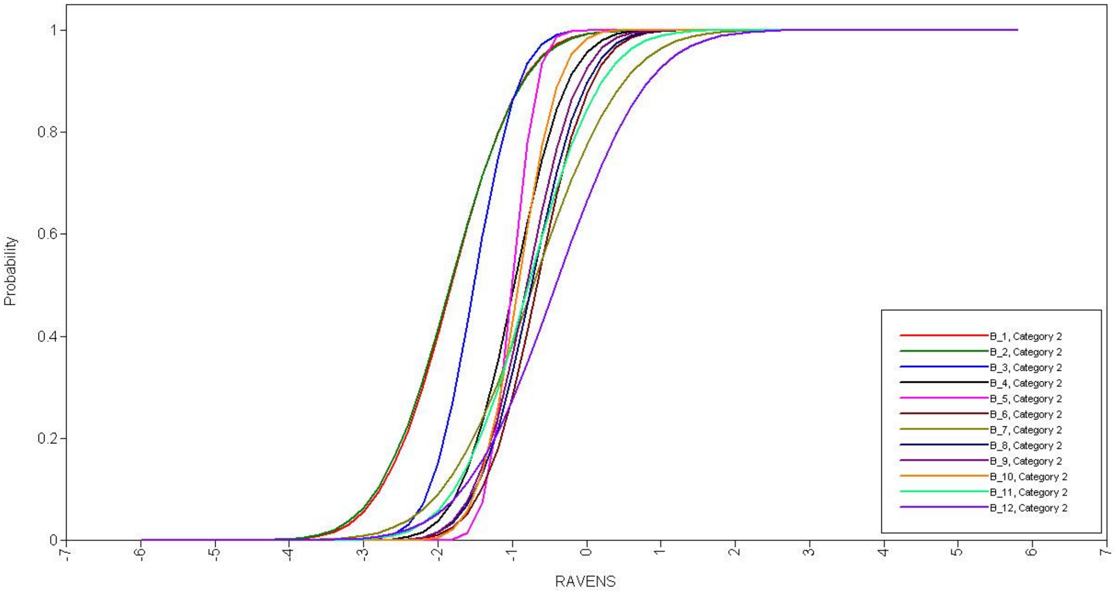


Figure 27: Female Item Characteristic Curves for the MG-CFA of Item Set B

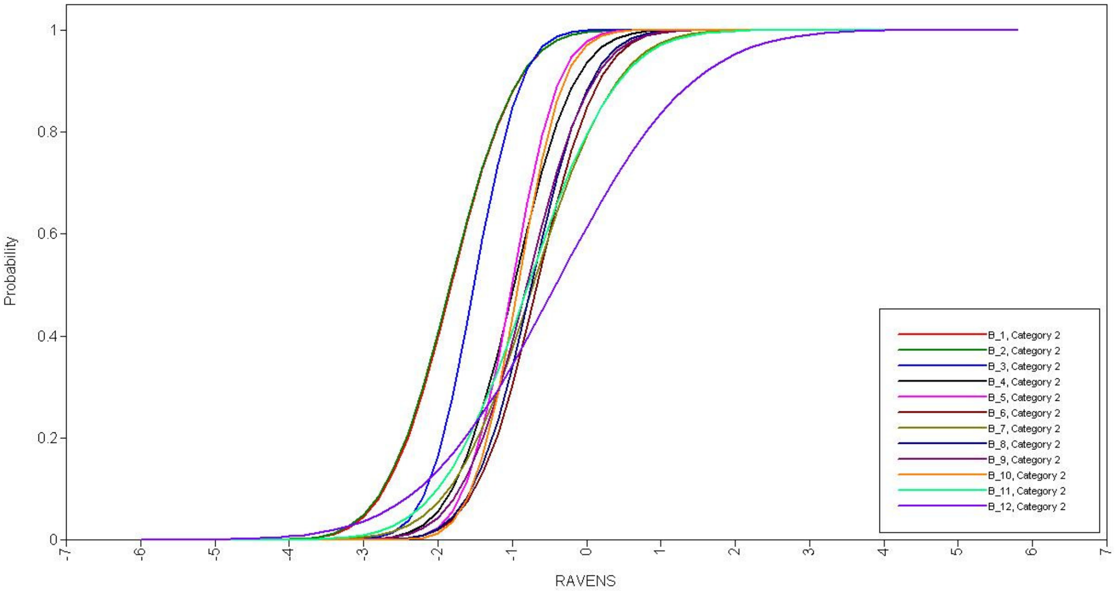


Figure 28: Male Item Characteristic Curves for the MG-CFA of Item Set C

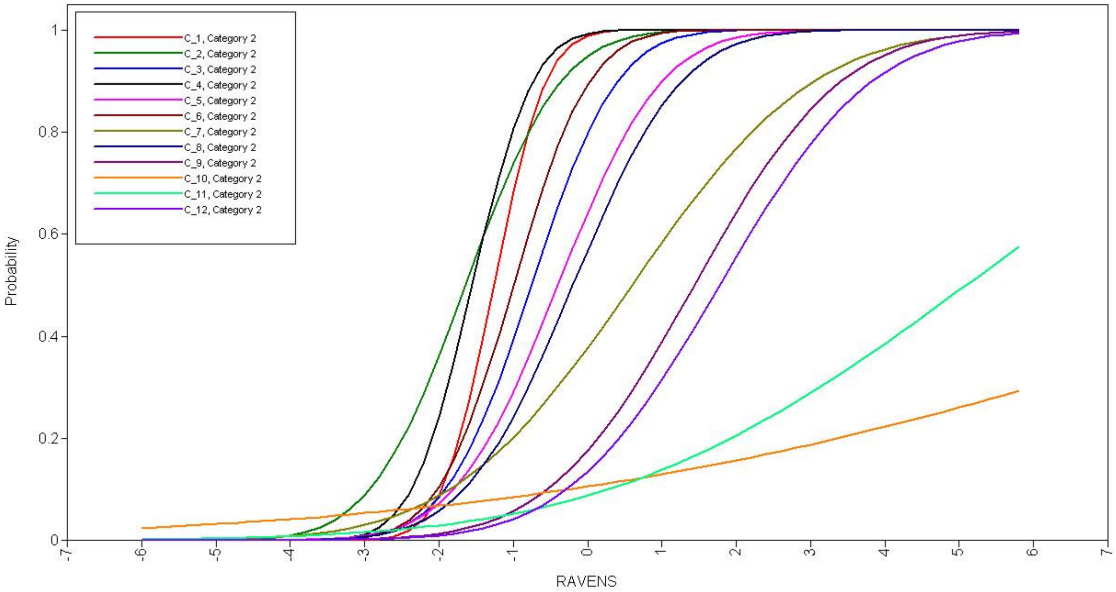


Figure 29: Female Item Characteristic Curves for the MG-CFA of Item Set C

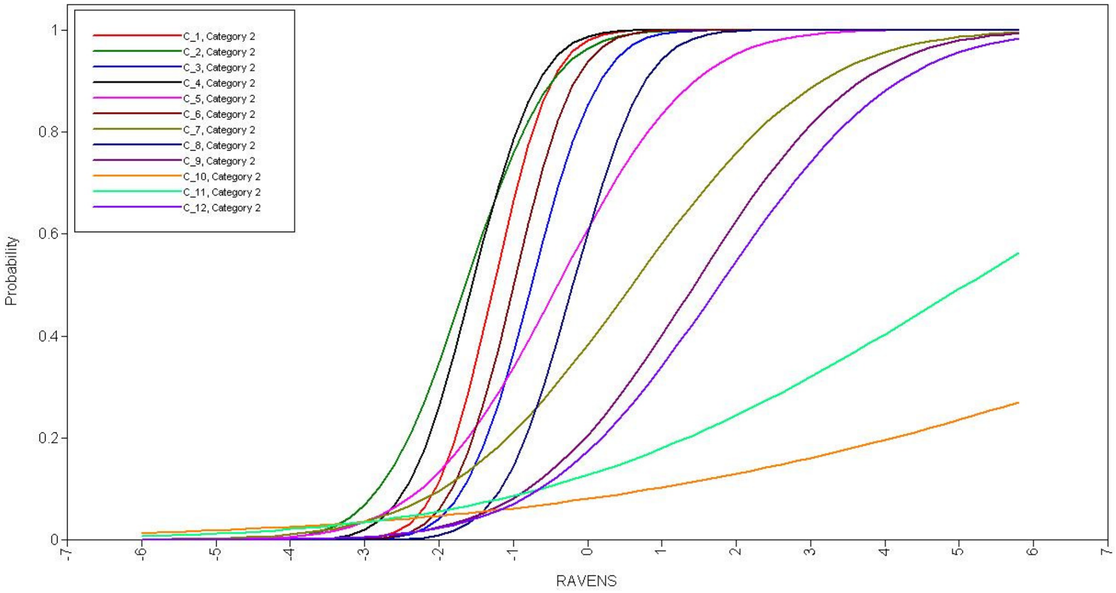


Figure 30: Male Item Characteristic Curves for the MG-CFA of Item Set D

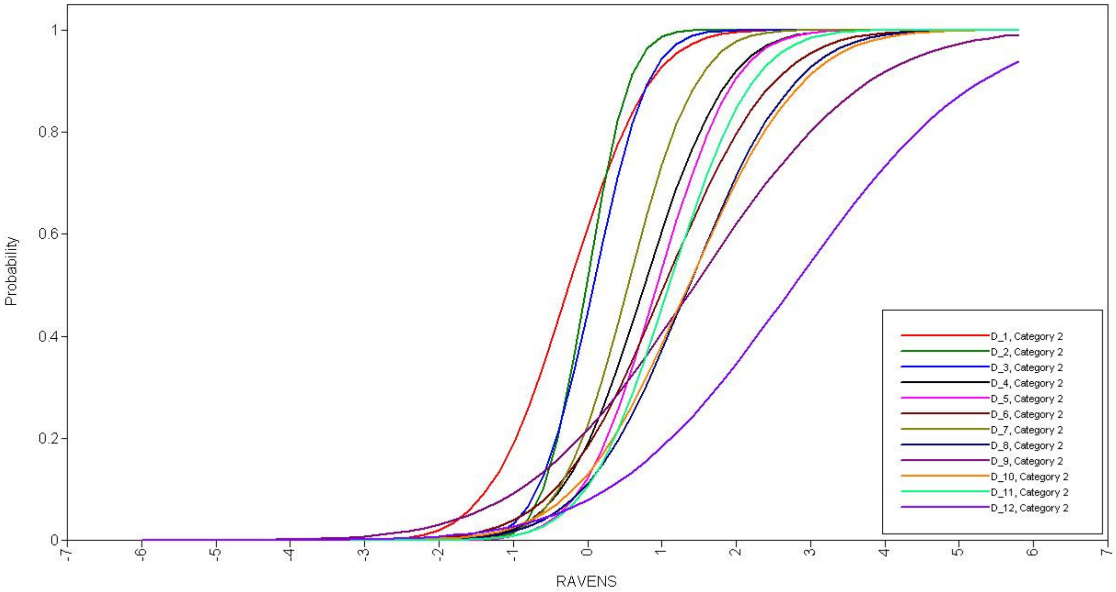


Figure 31: Female Item Characteristic Curves for the MG-CFA of Item Set D

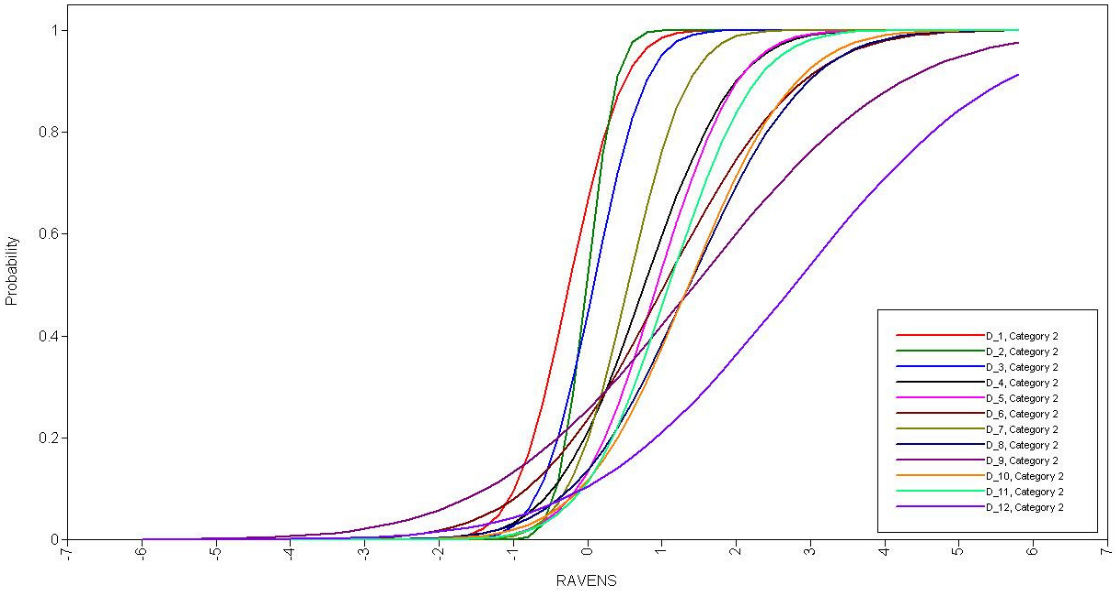


Figure 32: Male Item Characteristic Curves for the MG-CFA of Item Set E

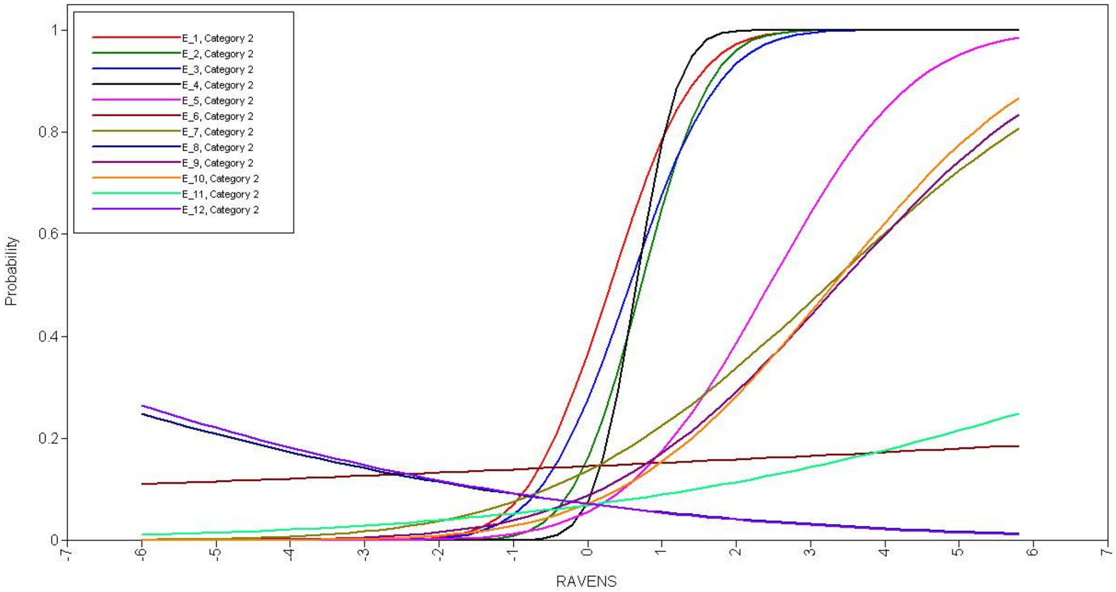
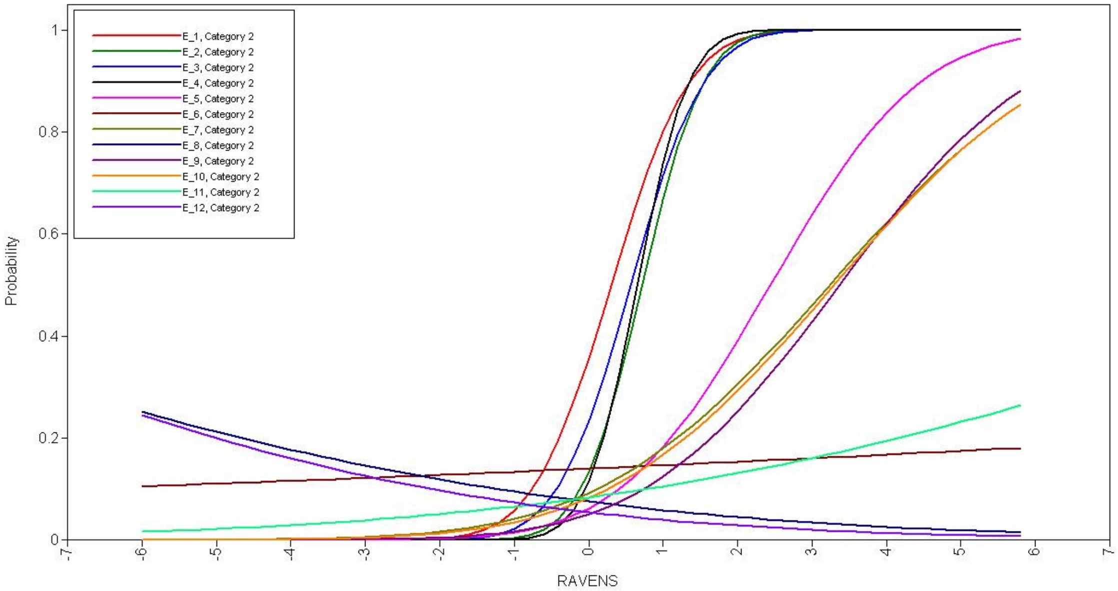


Figure 33: Female Item Characteristic Curves for the MG-CFA of Item Set E



Inspection of the ICCs suggests that, for both males and females, a number of the items are not fitting optimally to the current one-factor model, in particular items C10, C11, D12, and E5 through E12. This is evidenced by the item curves that exhibit flat or negative slopes and that do not follow the optimal 'S-shaped' curve. Examination of the standardised factor loadings reveals that 50 indicator factor loadings were statistically significant for males and 47 were statistically significant for females: $p < 0.01$. The latent factor explained significant variance: R^2_{Males} mean = 0.313, range = 0.018 to 0.796; $R^2_{Females}$ mean = 0.311, range = 0.003 to 0.678. However, there were a number of items with low communalities that did not reach statistical significance or salience with the suggested minimum threshold of 0.3 (McDonald, 1999).

For males, 3 of the 59 test items did not reach salience, but reached significance, while 6 items did not reach either salience or significance. For females, 5 of the 59 test items did not reach salience, but reached significance, while 7 items did not reach either salience or significance. Of these, 8 items non-significant and/or salient items were common between males and females (see Table 19). These findings are further supported by the ICCs in Figure 24 through Figure 33.

The next assessment of model fit will be with the inspection of the global fit indices, followed by a comparison of means and variability between boys and girls of the U.K. standardisation sample using MG-CFA methods.

6.2.1. MG-CFA Males and Females Simultaneously

In order to compare the performance of boys and girls on the SPM+ two separate correlation matrices were analysed simultaneously using MG-CFA. The models for males and females required different specification to optimise model fit. Figure 34 provides a diagram of the one-factor MG-CFA model applied for males. Figure 35 provides a diagram of the one-factor MG-CFA model applied for females.

To examine the equivalence of measurement (e.g., measurement invariance) and the structural features of the latent factors (e.g., population homogeneity), the factor loadings and indicator thresholds are constrained to equality across groups.

Figure 34: Diagram of the One-Factor MG-CFA for Males

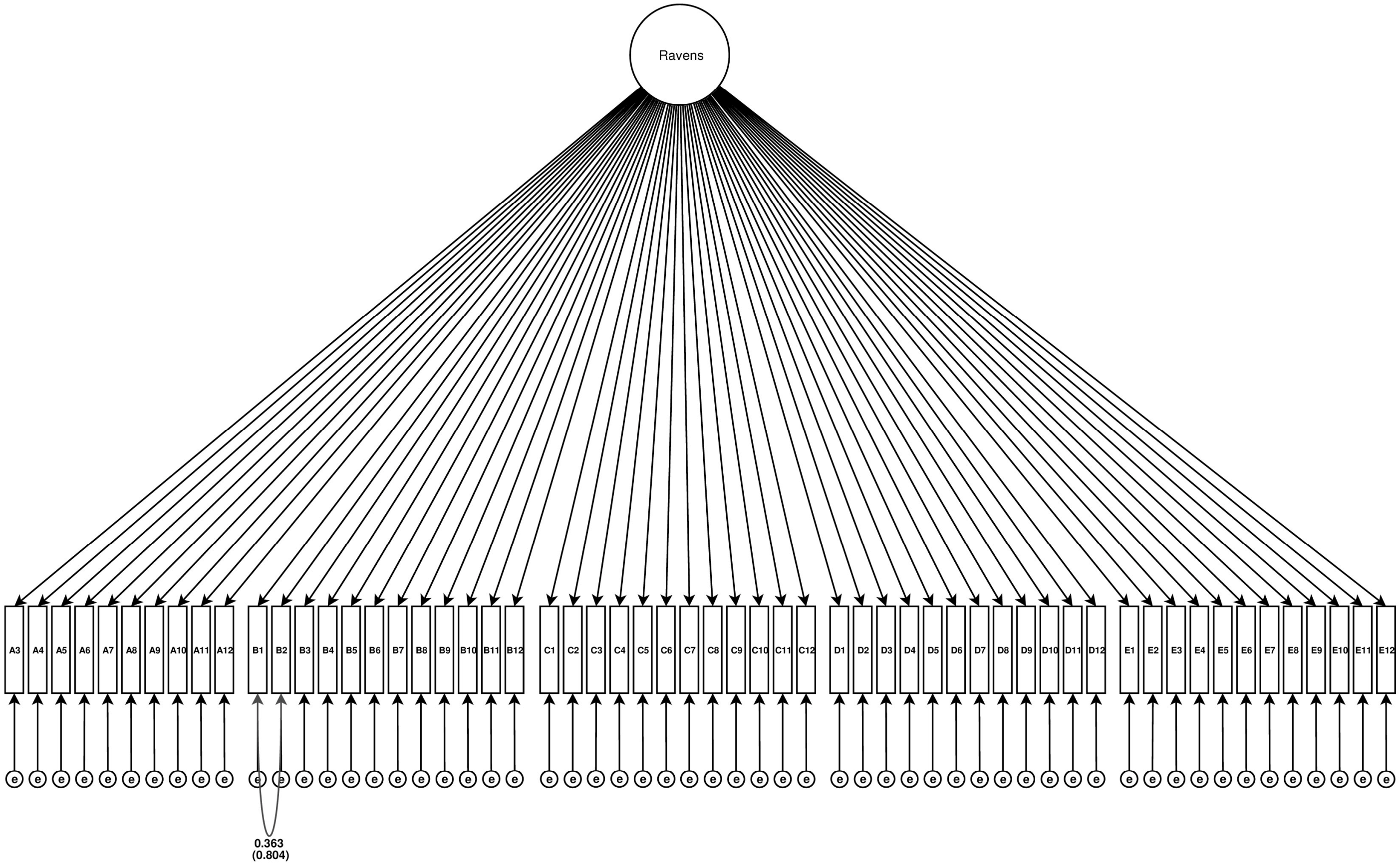
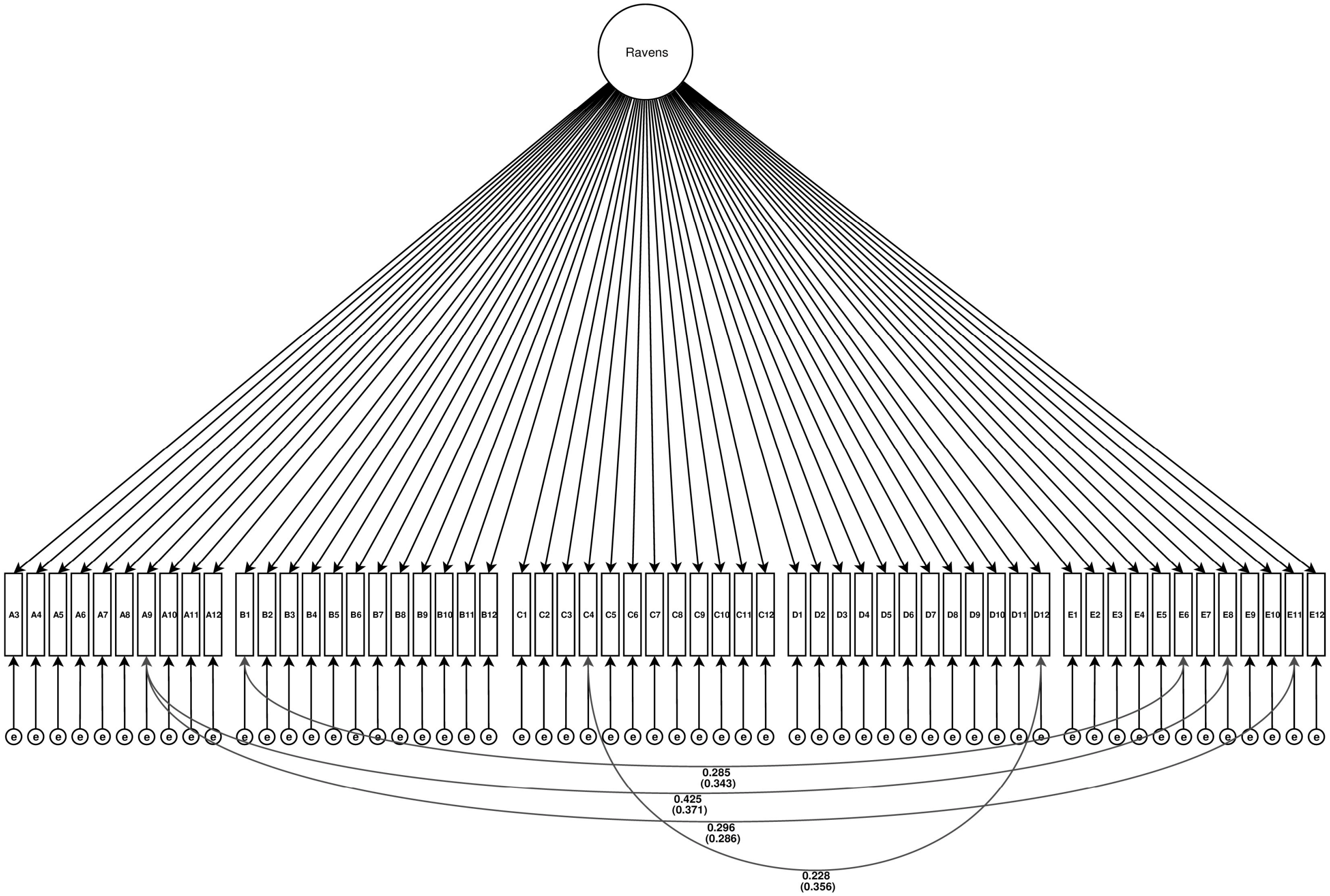


Figure 35: Diagram of the One-Factor MG-CFA for Females



By imposing such constraints systematically it allows us to rigorously test different aspects of the model. For example, by constraining the factor loadings to equality, the model is tested when the difference between 0 and 1 for the latent factor is the same for boys and girls. By constraining the indicator thresholds to equality, the model is tested when the value of the indicator is equivalent for boys and girls.

The measurement invariance model (M6-1) was over-identified with 352 *df*; χ^2 (352) = 584.275, *p* = 0.001. The statistics indicated good model fit: CFI = 0.901, TLI = 0.928, RMSEA = 0.038. The unstandardised and completely standardised parameter estimates, indicator thresholds and proportion of explained variance (R^2) of model M6-1 can be seen in Table 19.

Indicators that fail to significantly and saliently load are not considered to be well accounted for by the current latent factor model of general intelligence. This may be due to imposing equality constraints upon the model. Alternately, the differential covariance amongst the items may be attributable to the measurement approach, which will be investigated in a later section of this chapter.

To determine if the elimination of the non-significant and/or non-salient items from the CFA would significantly improve the model fit, the remaining 47 significant and salient items were loaded onto the one latent factor. The model (M6-1a) was over-identified with 293 *df*; χ^2 (293) = 542.999, $p < 0.001$, CFI = 0.904, TLI = 0.939, RMSEA = 0.043. A chi-square difference test was performed: $\chi^2_{\text{diff}}(59)$ = 41.276 (the critical value of chi-square at 59 *df* is 77.93, *p* = 0.05). From this it can conclude that the model is not significantly improved by the removal of the non-significant, non-salient items from model M6-1. Thus, it is suitable to retain these 12 items in subsequent analyses.

From the results of the one-factor models for both males and females, it has been concluded that all test items (A2 – E12) will be retained in future analyses for three reasons. First, according to the model fit indices, one latent factor of general intelligence is suitable for explaining the SPM+ data. Second, the removal of the non-significant, non-salient items did not result in a significant improvement in model fit. Finally, the SPM+ is a published test of intelligence that is used extensively in clinical and research settings, and all test items will be retained in the analyses for reasons of completeness, comparability to other published findings, and for generalisability to the population at large.

Inspection of the indicator thresholds of the 58 test items reveals that only item D2 did not have a significant threshold. An indicator threshold provides the value of the indicator when the latent factor is zero. When this value is significant it is indication that the value is significantly different than zero. Further, 31 of the 59 test thresholds were negative and 28 of the 59 were positive. Due to the binary nature of the data, these indicator thresholds can be interpreted in relation to item difficulty (Brown, 2006). In reference to Figure 5 in section 4.2, the values of the indicator difficulties are represented by the thresholds ranging from -1.993 for item A3 which is the easiest test item, to 1.574 for item E12 which is the most difficult item of the test.

To establish whether the indicator thresholds are significantly different for males and females, a chi-square difference test was conducted. First, the indicator thresholds were allowed to be freely estimated. With equality constraints imposed upon the factor loadings, the model (M6-2) was over-identified with 347 *df*; χ^2 (347) = 601.065, *p* = 0.001, CFI = 0.891, TLI = 0.920, RMSEA = 0.040. When this model was compared to model M6-1 with equality constraints imposed upon the factor loadings and indicator thresholds, there was not a significant degradation of fit to the model: $\chi^2_{\text{diff}}(5)$ = 6.790 (the critical value of chi-square at 5 *df* is 11.07, *p* = 0.05). From this it can be concluded that the indicator thresholds are not significantly different for boys and girls in the SPM+ sample.

To determine whether the factor loadings are significantly different for males and females, a further chi-square difference test was conducted. First, the factor loadings were allowed to be freely estimated. While the indicator thresholds were constrained to equality, the model (M6-3) was over-identified with 350 *df*; χ^2 (350) = 579.083, $p < 0.001$, CFI = 0.902, TLI = 0.928, RMSEA = 0.038. When this model was compared to model M6-1 with equality constraints imposed upon both the factor loadings and indicator thresholds, there was not a significant degradation of fit to the model: $\chi^2_{\text{diff}}(2)$ = 5.192 (the critical value of chi-square at 2 *df* is 5.99, *p* = 0.05). From the lack of significance in the chi-square difference test, it can be concluded that the factor loadings are not significantly different for boys and girls in the SPM+ sample.

To summarise, the overall evaluation of measurement invariance suggests that the factor loadings and item thresholds of the SPM+ latent factor are statistically equivalent for boys and girls. In other words, for each value of the latent factor of general intelligence, the observed values of each item are similar for boys and girls. It is now considered appropriate to proceed with a comparison of means and variability between the two groups of males and females.

6.3. Assessment of Sex Differences in Latent Mean and Variability

Having established that the SPM+ items measure the latent factor of general intelligence in the same way for boys and girls, it is now appropriate to extend the current one-factor MG-CFA (M6-1) in two ways. First, the variability of the latent factor will be constrained to equality for boys and girls (model M6-4) to assess differences in the latent factor mean and variance. Second, the boys' latent factor mean will be fixed to zero, to identify the mean structure component of the MG-CDFA. That is, the boys' latent factor mean serves as the reference and the girls' latent factor mean represents the difference between the two latent factors' means.

Power analyses demonstrate that the current sample size is sufficient to test models of sex differences on the Raven's SPM+. The probability of correctly rejecting null hypotheses of close, not close, and exact fit is 0.99 or higher with samples of 300 cases and models with 100 degrees of freedom (MacCallum et al., 1996).

The model was over-identified with 348 degrees of freedom: χ^2 (348) = 575.045, $p < 0.01$. The statistics indicated a good model fit: CFI = 0.903, TLI = 0.929, RMSEA = 0.038. The latent mean was 0.02 standard deviations higher for girls than for boys. This difference was not significant: z = 0.389, p = 0.698.

In relation to variance, the measurement invariance model (M6-1) solution showed that the overall variance was 0.282 for boys and 0.267 for girls. In order to assess whether the variance is significantly different between boys and girls, a constraint of equality is imposed on the variance of the model. The fit of the extended model was compared to that of the previous, measurement invariance model, M6-1. This analysis tests whether the SPM+ items drew on similar ranges of the latent factor of general intelligence for boys and girls. There was not a significant degradation to the model fit when variance of the latent factor of general intelligence was held to equality for boys and girls: χ^2_{diff} (4) = 9.23 (the critical value of chi-square at 4 *df* is 9.49, *p* = 0.05). From this, it can be concluded that the variance in general intelligence, as measured by the SPM+, is not significantly different between boys and girls.

6.4. Methods Factor

In each of the models presented thus far, a certain amount of variance has been unexplained by the one-factor solution. It is, therefore, prudent to attempt to account for this unexplained variance. This will be done by adding methods factors to the current set of one-factor models.

When a certain amount of the covariance among a set of items is due to the measurement approach rather than the substantive latent factor, it is known as a method effect or a method factor. A one-factor model for the Raven's SPM+ will now be expanded to include methods factors in order to account for the unexplained variance that was evident in previously described models from this chapter.

Table 21. Model Fit Statistics and Chi-square Difference Tests

Model	Chi-square statistic (χ^2)	Degrees of Freedom (*df*)	CFI	RMSEA	Chi-square Difference (χ^2_{diff})	Degrees of Freedom Difference (*df*$_{diff}$)	*p*
M6-1	584.275	352	0.901	0.038			
M6-1a	542.999	293	0.904	0.043	41.276[18]	59	NS
M6-2	601.065	347	0.891	0.040	6.790[19]	5	NS
M6-3	579.083	350	0.902	0.038	5.192[20]	2	NS
M6-4	575.045	348	0.903	0.038	9.23[21]	4	NS
M6-5a	318.475	203	0.904	0.034			
M6-5b	286.960	171	0.898	0.039			
M6-6	576.267	350	0.903	0.037			
M6-7a	568.912	347	0.905	0.037	1.974[22]	3	NS
M6-7b	568.537	347	0.905	0.037	2.349	3	NS
M6-8	566.339	346	0.906	0.037			

6.4.1. Model Specification and Measurement Invariance

Returning to model M6-1, 8% of the variance was unexplained by the latent factor model. As previously reported, this model was over-identified with 352 *df*; χ^2 (352) = 584.275, $p < 0.001$. The statistics indicated good model fit: CFI = 0.901, TLI = 0.928, RMSEA = 0.038. In specifying this model, the latent factor scores were saved to allow for further analyses of the residual variance. The latent factor scores were then regressed onto each indicator and the residual variances saved.

The residual variances were then subject to an exploratory factor analysis with geomin rotation to ascertain whether they grouped together in

[18] The critical value of chi-square at 59 *df* is 77.93, *p*=0.05.

[19] The critical value of the chi-square at 5 *df* is 11.07, p=0.05.

[20] The critical value of the chi-square at 2 *df* is 5.99, p=0.05.

[21] The critical value of the chi-square at 4 *df* is 9.49, p=0.05.

[22] The critical value of the chi-square at 3 *df* is 7.82, p=0.05.

any meaningful ways. Inspection of the EFA of the residuals revealed that 39 of the indicator residuals were either cross- or negatively loading, suggesting that were not suitable for inclusion in the methods factor. These residuals were subsequently removed from the analyses of the methods factors. The remaining 18 indicator residuals loaded onto two factors (Table 22).

Table 22: Results of the EFA of residual variance

Indicator	Residual Factor 1	Residual Factor 2
B1	0.570	
B3	0.561	
B2	0.497	
A9	0.496	
A6	0.491	
A8	0.356	
B4	0.333	
A4	0.329	
A7	0.320	
B9		0.560
B10		0.538
B8		0.492
B11		0.427
C1		0.345
B6		0.313
C6		0.292
B12		0.254
C3		0.183

Inspection of the residual factors reveal that the indicators are located at the beginning of the SPM+: item sets A, B, and the first half of set C. According to the results from the Rasch models presented in Chapter 5 these indicators are the easiest of the test. In consultation with the existing literature of the strategies used to solve items of the SPM (Deshon, Chan, & Weissbein, 1995; Lynn, Allik, & Irwing, 2004; van der Ven & Ellis, 2000), it could be suggested that the residual factors of the current model comprise items influenced by Gestalt and Visuospatial solving strategies (Residual Factors 1 and 2 respectively).

The results from the EFA of the residuals informed the specification of a CFA with methods factors. The 58 SPM+ test items were loaded onto one latent factor, with 18 items loaded onto two secondary methods factors. This model was analysed for males and females separately. The model for females (M6-5a) was over-identified with 203 *df*; χ^2 (203) = 318.475, $p < 0.001$. The statistics indicated good model fit: CFI = 0.904, TLI = 0.939, RMSEA = 0.034. For males, the model (M6-5b) was also over-identified with 171 *df*; χ^2 (171) = 286.96, $p < 0.001$. The statistics indicated good model fit: CFI = 0.898, TLI = 0.924, RMSEA = 0.039. Inspection of the modification indices suggests that there are no points of ill fit within either of the models.

The models M6-5a and M6-5a were further expanded, by analysing the male and female input matrices simultaneously. The model (M6-6) was over-identified with 350 *df*; χ^2 (350) = 576.267, $p < 0.001$. The statistics indicated good model fit: CFI = 0.903, TLI = 0.929, RMSEA = 0.037. No points of ill fit were identified within the model by the modification indices. Details of the minimum, maximum, and average factor loadings, indicator thresholds and proportion of explained variance (R^2) are available in Table 23.

It is now appropriate to extend the MG-CFA$_{MF}$ in order to assess the mean differences in the methods factors between males and females. First, each method factor will be assessed separately using a chi-square difference test. The variability of the first method factor was held to equality. The model (M6-7a) was over-identified with 3 *df*; χ^2 (3) = 568.912, $p < 0.001$. The statistics indicated good model fit: CFI = 0.905, TLI = 0.930, RMSEA = 0.037. A chi-square difference test was performed χ^2_{diff} (3) = 1.974 (the critical value of the chi-square at 3 *df* is 7.82, p = 0.05). The variance prior to equality constraint was 0.069 for males, and 0.179 for females, and after the constraint was 0.270 for males and 0.219 for females.

Table 23: Minimum, Maximum and Average Factor Loadings, Indicator Thresholds and R-squares for a 1-factor model with a methods factor for males and females

		Male	Female
Factor Loadings	**Min**	-0.030	-0.031
	Max	0.903	0.846
	Average	0.446	0.318
Indicator Thresholds	**Min**	-2.051	-2.279
	Max	1.594	1.510
	Average	-1.278	-0.305
R^2	**Min**	0.001	0.001
	Max	0.852	0.852
	Average	0.109	0.200

Next, the variability of the second method factor was held to equality. The model (M6-7b) was over-identified with 347 *df*; χ^2 (347) = 568.537, *p < 0.001*. The statistics indicated good model fit: CFI = 0.905, TLI = 0.930, RMSEA = 0.037. A chi-square difference test was performed χ^2_{diff} (3) = 2.349 (the critical value of the chi-square at 3 *df* is 7.82, *p* = 0.05). The variance of the second factor prior to equality constraint was 0.225 for males and 0.235 for females and after the constraint was 0.265 for both males and females.

Finally, to assess the mean differences of the methods factors between males and females, the variability of the factors was constrained to equality to assess mean differences in the methods factors. The model (M6-8) was over-identified with 346 *df*; χ^2 (346) = 566.339, *p* =.001. The statistics indicated good model fit: CFI=0.906, TLI=0.930, RMSEA=0.037. The method factor mean for first methods factor was 0.549 standard deviations greater for females than for males, but this difference was not significant: z = 1.268, p = 0.205. The mean for second methods factor was 0.384 standard deviations higher for females than for males. This difference was significant: z = 2.224, p = 0.026.

6.5. SUMMARY

This chapter has provided an assessment of a one-factor model of general intelligence for the Raven's SPM+ data from the U.K. standardisation sample using two forms of confirmatory factor analysis: CFA for males and females separately and MG-CFA factor for males and females simultaneously. Further, methods factors were added to the MG-CFA model to assess if any of the residual variance could be attributed to something other than the latent construct of general intelligence. In doing so, three of the research objectives of this dissertation were satisfied.

First, it is evident from the goodness-of-fit statistics that the Raven's SPM+ is adequately represented using a one-factor model. Further, the least constrained, measurement invariance model (M6-1) provides evidence that there is no bias present at the level of factor loadings or item thresholds. That is, the metric for boys and girls at each value of the latent factor is equivalent, or in other words, items on the SPM+ are measuring general intelligence in the same way for both boys and girls.

Having determined this, it was appropriate to assess the second objective: sex differences in the latent factor mean and variance. When the variance of the latent factors was constrained to equality for boys and girls, girls in the standardisation sample achieved mean scores 0.02 standard deviations higher than boys, but not significantly so. Further, there was no significant degradation to the model, indicating that the variance in the latent factor is not significantly different between boys and girls.

Finally, when the MG-CFA model was expanded to include the methods factors, the measurement invariance model indicates that the metric of the methods factors is equivalent for boys and girls. In other words, the influence of the measurement approach is the same for boys and girls. When each of the methods factors was isolated in succession, it was revealed that there was no significant difference in the mean of the first methods factor (or what might be thought of as Gestalt items). There was a significant difference between boys and girls on the second methods factor (or what have been proposed by some as Visuospatial items; Carpenter, Just, & Shell, 1990; Lynn, Allik, & Irwing, 2004; van der Ven & Ellis, 2000). This may be suggesting that once general intelligence is accounted for by the latent factor, males are significantly negatively affected by the visuospatial element of the items B6, B8, B9, B10, B11, B12, C1, and C6.

It is important to consider that these results are from the complete U.K. standardisation sample comprising individuals ranging in age from 7 years 0 months to 18 years 11 months. Across such a large age range (11 years 11 months), the factor structure of the SPM+ may not be stable.

In accordance with this, Lynn (1994) has proposed the "Developmental Theory of Sex Differences" which states that girls mature earlier than boys, resulting in a cognitive advantage over boys between the ages of 10-13 years. By 15 years of age, Lynn suggests that there is a developmental deceleration for girls while boys continue to develop, resulting in a cognitive male advantage from 15-16 years of age and onward. In order to account for the impact of age in the analyses of sex differences on the SPM+, MG-CFA techniques will now be used with four groups (younger girls, younger boys, older girls, older boys) in Chapter 7 to ascertain whether there are sex differences in general intelligence that emerge at different ages along the developmental continuum.

— 7 —

RESULTS CHAPTER 3: ASSESSMENT OF GROUP DIFFERENCES IN MEAN AND VARIABILITY USING MULTIPLE GROUPS CONFIRMATORY FACTOR ANALYSIS OF YOUNGER & OLDER PARTICIPANTS OF THE SPM+

7.1. INTRODUCTION

The third objective of this dissertation will be addressed in this chapter: Are there sex differences in younger or older participants of the SPM+? The fourth dissertation objective of the association of methods effects will also be investigated further in this chapter.

In previous chapters, the literature reporting inconsistent findings of sex differences in general intelligence has been discussed (Jackson & Rushton, 2006; Lynn, 1994; Lynn, 1998; Rushton & Cvorovic, 2009). One suggested explanation of the inconsistent findings relates to the lack of psychometrically sound assessment of item bias and measurement errors prior to the assessment of group differences. This was addressed in Chapter 6 where the suitability of a one-factor solution for two groups (males and females) was presented, having first assessed the measurement properties of the model prior to the evaluation of male-female differences.

Another way of thinking about sex differences is within the context of a developmental trajectory: boys and girls mature at different rates, and differences in intelligence may be influenced by the variation in development of the two sexes. For generations, it has been suggested that girls mature earlier than boys with respect to the development of a number of physical and cognitive characteristics (Hohm et al., 2007; Nature, 1923; de Onis et al., 2007). According to Lynn (1999, 2004), failure to account for differences in maturation between boys and girls may be masking true sex differences in general intelligence in a number of studies of children and young adults.

According to the maturational differences, Lynn contends that a female advantage would be apparent before puberty, while a male advantage would begin to emerge after puberty in late adolescence or early adulthood. His "*Developmental Theory of Sex Differences*" (Lynn, 2002) proposed that girls mature earlier than boys, both cognitively and physically, and tend to have a cognitive advantage over males of about 1 IQ point between eight to 15 years. By 15 years of age, however, there is a developmental deceleration for females while boys continue to develop. Lynn claims that this results in a male advantage of approximately 2.4 IQ points from approximately 16 years of age, an advantage that is maintained throughout adulthood (Lynn, 2002; Lynn, Allik, & Must, 2000). Evidence of Lynn's Developmental Theory is available from a number of published studies of sex differences on the Raven's SPM. However, the results do not match the developmental pattern or direction of difference proposed by Lynn, resulting in a literature that is somewhat contradictory and inconclusive.

A study conducted by Abdel-Khalek and Lynn (2006) assessed sex differences in eight to 15 year olds in Kuwait determined that girls outperformed boys between eight to 12 years (d = -0.06 to -0.27). At this point, a small non-significant male advantage began to emerge in boys between 13 to 15 years of age (d = 0.01 to 0.06). A further study by Lynn, Backhoff, and Contreras-Niño (2004) identified a developmental trend in the scores of a large sample of seven to 10 year old children in Mexico. A slight male advantage was apparent at seven years but this decreased to a point of non-significant female advantage at 10 years.

A meta-analysis conducted by Lynn and Irwing (2004) concludes that boys obtain slightly higher means than girls from six to nine years of age, but not significantly so (*d* = 0.01 to 0.10). From 10 to 13 years, a higher non-significant mean emerges for females (*d* = -0.06 to 0.05). At 14 years of age, a male advantage emerges (*d* = 0.08) which, at 15 years, becomes significant and increases in effect size to (*d* = 0.10). By 18 years of age, the significant difference increases in size to 0.16 *d*.

The conflicting research conclusions are also evidenced in samples of older children and adults. In a sample of 12 to 18 year olds in Estonia, (Lynn, Allik, & Irwing, 2004) determined that girls performed better than boys from 12 to 13 years of age (*d* = -0.384)[23], there were no differences between the sexes between 14 to 16 years of age (*d* = -.033), but a male advantage emerged at 17 years of age (*d* = 0.193). Using a standardisation sample of the SPM in Estonia, (Lynn, Allik, Pullmann, & Laidra, 2004) conclude a female advantage among 12 to 15 year old (*d* = -0.03 to -0.54), and a male advantage between 16 to 18 year olds (*d* = 0.04 to 0.80). In a further study, Rushton and Cvorovic (2009) claimed that there are no mean differences between males and females according to their study of a sample of 17 to 65 year olds.

While there is available evidence from the literature that age is involved in mean score differences between males and females along the developmental continuum, the conflicting results do not lend themselves towards supporting a cohesive Developmental Theory of Sex Differences. As discussed previously, a number of these studies have made use of opportunity samples or statistical analyses that are not sensitive to detecting true group differences. As such, there is an identified need in the literature for a study that accounts for age in the assessment of sex differences in general intelligence using a representative sample and sound, statistical methodology.

In the current dissertation, the age-range of the participants of the U.K. standardisation is from seven to 18 years of age. The age range of participants from the standardisation sample spans the periods of developmental change proposed by Lynn's Developmental Theory of Sex

[23] As previously indicated the negative effect size indicates a female advantage.

Differences. During this broad age range of 11 years and 11 months, it is conceivable that age of the participant could be a contributing element in their performance on the SPM+, and could subsequently inform overall conclusions drawn.

For this reason, it is important to examine the issue of sex differences in general intelligence while accounting for the age of the participants of the SPM+. To do so, Multiple-Group Confirmatory Factor Analysis (MG-CFA) techniques will again be employed in order to compare the performance of boys and girls in younger (7 years 0 months to 14 years 11 months) and older (15 years 0 months to 18 years 11 months) age groups. The point at which the groups are divided is guided by Lynn's theory. These four groups (younger boys, younger girls, older boys, and older girls) will be compared with respect to the equivalence of measurement characteristics (i.e., factor loadings, indicator thresholds, residual variances) and structural features (i.e., factor variances, factor covariances, latent means) to assess population homogeneity.

As discussed in Chapters 4 and 6, it is of utmost importance to ensure that the measurement properties are equivalent in each group in order to make meaningful comparisons between multiple groups (Brown, 2006; van Der Sluis et al., 2008). Further, increasing the complexity of the MG-CFA techniques allows for the assessment of item difficulty, discrimination, and method factor effects separately in younger and older age groups. This allows greater specificity of identifying where, if any, sex differences in general intelligence exist.

It has already been established that a one-factor Confirmatory Factor Analysis model of general intelligence is suitable for use with the current standardisation sample (sections 5.5 and 6.2). The one-factor model will be elaborated by using four age-group input correlation matrices. The measurement properties will be assessed for equivalence for younger girls, younger boys, older girls and older boys in a one-factor model (section 7.2). The latent factor means and variability will be compared between: 1) younger and older participants; 2) younger boys and girls; and 3) older boys and girls (section 7.3). Finally, methods effects will be tested for the males and females in the younger and older age groups in an attempt to account for additional unexplained variance in the models (section 7.4).

7.2. MODEL SPECIFICATION & MEASUREMENT INVARIANCE

A suitable one factor solution has already been established to model the SPM+ for the whole standardisation sample (Model M5-1, Chapter 5), as well as for males and females separately (Model M6-1, Chapter 6). To determine how well the items on the SPM+ can be explained by one underlying latent trait of general intelligence for four age-related groups of the standardisation sample (younger males, younger females, older males and older females), a unidimensional confirmatory factor analysis was initially conducted.

The 58 SPM+ test items were loaded onto one latent factor and analysed in four separate input correlation matrices, one for each of the following groups: younger males (n = 323), younger females (n = 340), older males (n = 114) and older females (n = 149). Preliminary analysis revealed that all of the older females correctly answered item A6, resulting in a lack of variance. This item was removed from further analyses for older females. Similarly, all of the older males answered item B3 correctly, which was subsequently removed from further analyses due to a lack of variance.

There is good fit for a one-factor model for the younger male and female groups, with adequate fit for the older males and females. Details of the model fit statistics for each group are available in Table 24.

Table 24: Model fit statistics for a one-factor model of the SPM+ for young males, young females, older males and older females

Model Fit Statistics	Young Males (n = 323)	Young Females (n = 340)	Older Males (n = 114)	Older Females (n = 149)
Chi-Square (*df*)	245.989 (138)	254.310 (162)	68.797 (48)	79.670 (48)
CFI	0.856	0.865	0.794	0.684
TLI	0.885	0.903	0.790	0.704
RMSEA	0.049	0.041	0.062	0.067

Power analyses demonstrate that the current sample size for the younger sample is sufficient to test models of sex differences on the Raven's SPM+. The probability of correctly rejecting null hypotheses of close, not close, and exact fit is 0.99 or higher with samples of 300 cases and models with 100 degrees of freedom. For the older participants, the current sample size is less than optimal for testing models of sex differences. The probability of correctly rejecting the null hypotheses of close, not close, and exact fit is between 0.261 to 0.424 with 50 degrees of freedom (MacCallum et al., 1996). In order to achieve a minimum power of 0.80, the sample size of the older sample would need to be increased by approximately 250 participants.

Inspection of the standardised residuals and modification indices indicate that there are no localised points of ill fit within the model. Details of the minimum, maximum, and average factor loadings, indicator thresholds and proportion of explained variance (R^2) for each of the four groups are available in Table 25.

Table 25: Minimum, Maximum and Average Factor Loadings, Indicator Thresholds, Proportion of Variance Explained (R^2) and Factor Variance for 1-factor models for Younger Males, Younger Females, Older Males and Older Females

		Younger Males (n=323)	Younger Females (n=340)	Older Males (n=114)	Older Females (n=149)
Factor Loadings	**Min**	-0.164	-0.238	-0.145	-0.152
	Max	0.894	0.873	0.864	0.837
	Average	0.445	0.426	0.460	0.444
Indicator Thresholds	**Min**	-1.842	-2.067	-1.833	-2.164
	Max	1.689	1.389	1.215	1.903
	Average	0.055	0.040	0.214	0.113
R^2	**Min**	0.000	0.000	0.000	0.000
	Max	0.800	0.763	0.746	0.700
	Average	0.259	0.246	0.264	0.256
Factor Variance	**Unstand.**	0.372	0.345	0.560	0.451
	Standard.	1.000	1.000	1.000	1.000

In order to compare the performance of the four age groups, the four input matrices were analysed simultaneously (younger boys, younger girls, older boys, and older girls). 56 test items were loaded onto one latent factor of general intelligence. Figure 36 provides a diagram of the one-factor MG-CFA model applied simultaneously for young males, younger females, older males and older females.

To examine whether the measurement properties were equivalent across all four groups, the factor loadings and indicator thresholds were constrained to equality. The measurement invariance model (M7-1) was over-identified with 297 *df*; χ^2 (297) = 486.046, $p < 0.001$. The statistics indicated good model fit: CFI = 0.839, TLI = 0.847, RMSEA =0.052. The unstandardised and completely standardised parameter estimates, indicator thresholds and proportion of explained variance (R^2) of model

M7-1 can be seen in Table 26 for younger males and females and in Table 27 for older males and females.

Figure 36: Diagram of the Four-Group One-Factor MG-CFA for Younger Males, Younger Females, Older Males and Older Females

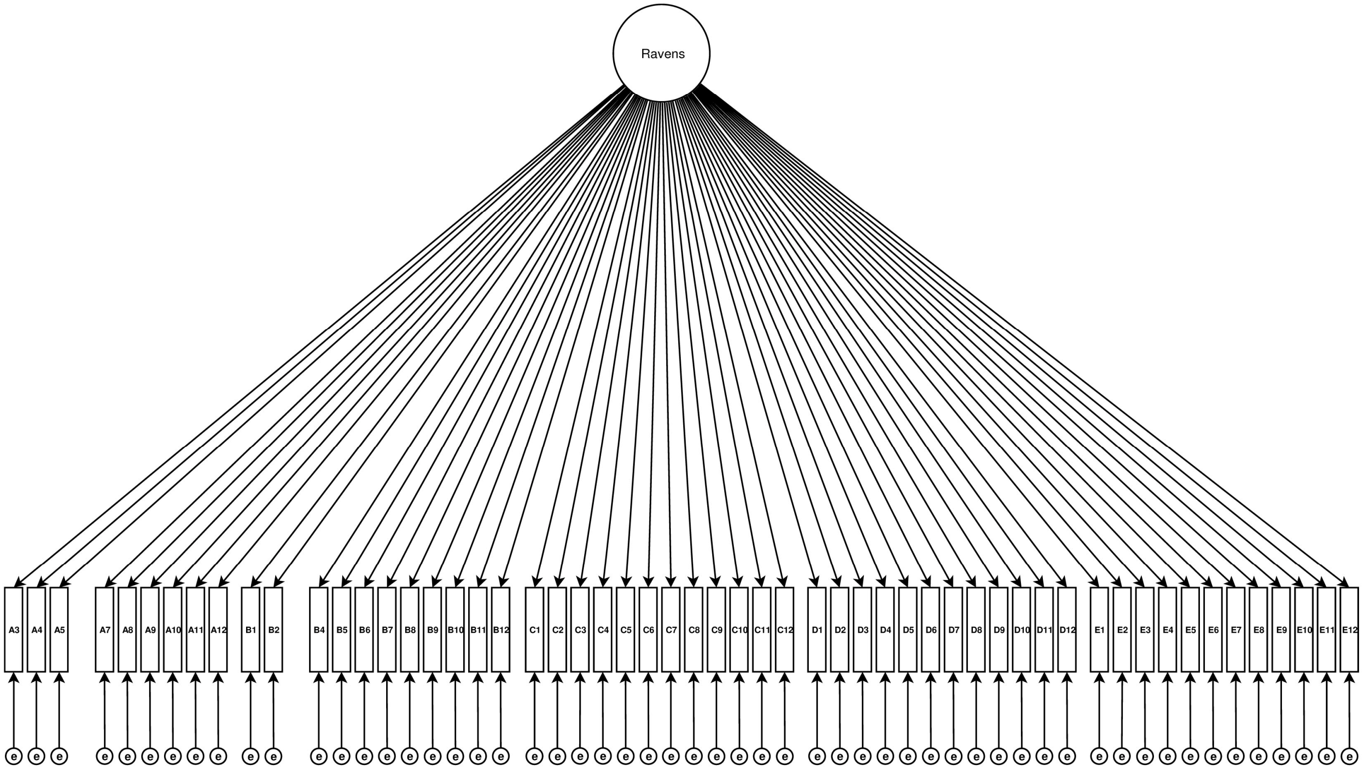


Table 26: Unstandardised Parameter Estimates, Indicator Thresholds, and Proportion of Variance Explained (R^2) for Younger Males and Younger Females

	Male		Female		Male		Female		Male		Female		Male		Female		Male		Female	
Indicator	Unstand. Factor Loading	*p*	Unstand. Factor Loading	*p*	Standardised Factor Loading	*p*	Standardised Factor Loading	*p*	Unstand. Indicator Threshold	*p*	Unstand. Indicator Threshold	*p*	Standardised Indicator Threshold	*p*	Standardised Indicator Threshold	*p*	Prop. of Variance Explained (R^2)	*p*	Prop. of Variance Explained (R^2)	*p*
A3	1.000	0.000	1.000	0.000	0.610	0.000	0.659	0.000	-1.842	0.000	-1.842	0.000	-1.842	0.000	-2.067	0.000	0.372	0.000	0.434	0.004
A4	1.032	0.000	1.032	0.000	0.630	0.000	0.568	0.000	-1.944	0.000	-1.944	0.000	-1.944	0.000	-1.822	0.000	0.396	0.000	0.322	0.011
A5	0.932	0.000	0.932	0.000	0.569	0.000	0.665	0.000	-1.601	0.000	-1.601	0.000	-1.601	0.000	-1.944	0.000	0.324	0.001	0.442	0.000
A6	0.904	0.000	0.904	0.000	0.552	0.000	0.428	0.000	-1.282	0.000	-1.282	0.000	-1.282	0.000	-1.035	0.000	0.304	0.000	0.183	0.001
A7	1.137	0.000	1.137	0.000	0.694	0.000	0.565	0.000	-1.024	0.000	-1.024	0.000	-1.024	0.000	-0.866	0.000	0.482	0.000	0.319	0.000
A8	1.187	0.000	1.187	0.000	0.725	0.000	0.518	0.000	-1.528	0.000	-1.528	0.000	-1.528	0.000	-1.134	0.000	0.525	0.000	0.268	0.001
A9	1.164	0.000	1.164	0.000	0.710	0.000	0.564	0.000	-1.091	0.000	-1.091	0.000	-1.091	0.000	-0.900	0.000	0.504	0.000	0.318	0.000
A10	0.922	0.000	0.922	0.000	0.562	0.000	0.439	0.000	-0.595	0.000	-0.595	0.000	-0.595	0.000	-0.484	0.000	0.316	0.000	0.193	0.000
A11	0.530	0.000	0.530	0.000	0.324	0.000	0.354	0.000	-0.166	0.002	-0.166	0.002	-0.166	0.002	-0.189	0.002	0.105	0.018	0.125	0.006
A12	1.173	0.000	1.173	0.000	0.716	0.000	0.674	0.000	-1.897	0.000	-1.897	0.000	-1.897	0.000	-1.857	0.000	0.513	0.000	0.455	0.000
B1	0.947	0.000	0.947	0.000	0.578	0.000	0.710	0.000	-1.490	0.000	-1.490	0.000	-1.490	0.000	-1.904	0.000	0.334	0.000	0.504	0.000
B2	1.301	0.000	1.301	0.000	0.794	0.000	0.738	0.000	-1.014	0.000	-1.014	0.000	-1.014	0.000	-0.980	0.000	0.630	0.000	0.545	0.000
B3	1.466	0.000	1.466	0.000	0.894	0.000	0.785	0.000	-1.258	0.000	-1.258	0.000	-1.258	0.000	-1.147	0.000	0.800	0.000	0.616	0.000
B4	1.194	0.000	1.194	0.000	0.728	0.000	0.744	0.000	-0.687	0.000	-0.687	0.000	-0.687	0.000	-0.730	0.000	0.531	0.000	0.554	0.000
B5	0.925	0.000	0.925	0.000	0.564	0.000	0.655	0.000	-0.568	0.000	-0.568	0.000	-0.568	0.000	-0.685	0.000	0.318	0.000	0.429	0.000
B6	1.220	0.000	1.220	0.000	0.744	0.000	0.714	0.000	-0.749	0.000	-0.749	0.000	-0.749	0.000	-0.746	0.000	0.554	0.000	0.509	0.000
B7	1.149	0.000	1.149	0.000	0.701	0.000	0.691	0.000	-0.873	0.000	-0.873	0.000	-0.873	0.000	-0.895	0.000	0.492	0.000	0.478	0.000
B8	1.315	0.000	1.315	0.000	0.803	0.000	0.873	0.000	-1.022	0.000	-1.022	0.000	-1.022	0.000	-1.156	0.000	0.645	0.000	0.763	0.000
B9	1.060	0.000	1.060	0.000	0.647	0.000	0.605	0.000	-0.770	0.000	-0.770	0.000	-0.770	0.000	-0.748	0.000	0.419	0.000	0.366	0.000
B10	0.957	0.000	0.957	0.000	0.584	0.000	0.478	0.000	-0.273	0.000	-0.273	0.000	-0.273	0.000	-0.232	0.000	0.341	0.000	0.228	0.000
B11	1.296	0.000	1.296	0.000	0.791	0.000	0.720	0.000	-1.184	0.000	-1.184	0.000	-1.184	0.000	-1.119	0.000	0.626	0.000	0.518	0.000
B12	0.948	0.000	0.948	0.000	0.578	0.000	0.633	0.000	-1.135	0.000	-1.135	0.000	-1.135	0.000	-1.292	0.000	0.334	0.000	0.401	0.000
C1	0.969	0.000	0.969	0.000	0.591	0.000	0.632	0.000	-0.507	0.000	-0.507	0.000	-0.507	0.000	-0.564	0.000	0.350	0.000	0.400	0.000
C2	1.144	0.000	1.144	0.000	0.698	0.000	0.718	0.000	-1.308	0.000	-1.308	0.000	-1.308	0.000	-1.398	0.000	0.487	0.000	0.515	0.000
C3	0.999	0.000	0.999	0.000	0.610	0.000	0.453	0.000	-0.051	0.413	-0.051	0.413	-0.051	0.413	-0.039	0.421	0.372	0.000	0.205	0.000
C4	1.034	0.000	1.034	0.000	0.631	0.000	0.708	0.000	-0.753	0.000	-0.753	0.000	-0.753	0.000	-0.877	0.000	0.399	0.000	0.501	0.000
C5	0.636	0.000	0.636	0.000	0.388	0.000	0.296	0.000	0.497	0.000	0.497	0.000	0.497	0.000	0.394	0.000	0.151	0.000	0.088	0.016
C6	0.916	0.000	0.916	0.000	0.559	0.000	0.540	0.000	0.064	0.277	0.064	0.277	0.064	0.277	0.064	0.268	0.313	0.000	0.292	0.000
C7	0.646	0.000	0.646	0.000	0.394	0.000	0.410	0.000	0.970	0.000	0.970	0.000	0.970	0.000	1.047	0.000	0.155	0.001	0.168	0.000

*Item factor loading did not reach the minimum required threshold of 0.3 (McDonald, 1999).
**Item factor loading did not reach statistical significance.

	Male		Female		Male		Female		Male		Female		Male		Female		Male		Female	
Indicator	Unstand. Factor Loading	*p*	Unstand. Factor Loading	*p*	Standardised Factor Loading	*p*	Standardised Factor Loading	*p*	Unstand. Indicator Threshold	*p*	Unstand. Indicator Threshold	*p*	Standardised Indicator Threshold	*p*	Standardised Indicator Threshold	*p*	Prop. of Variance Explained (R^2)	*p*	Prop. of Variance Explained (R^2)	*p*
C8	0.167	0.021	0.167	0.021	0.102*	0.021	0.119*	0.015	1.222	0.000	1.222	0.000	1.222	0.000	1.478	0.000	0.010	0.249	0.014	0.225
C9	0.484	0.000	-0.437	0.028	0.296	0.000	-0.238*	0.025	1.319	0.000	1.319	0.000	1.319	0.000	1.222	0.000	0.087	0.002	0.056	0.261
C10	0.653	0.000	0.653	0.000	0.399	0.000	0.319	0.000	1.301	0.000	1.301	0.000	1.301	0.000	1.080	0.000	0.159	0.000	0.101	0.002
C11	0.805	0.000	0.805	0.000	0.491	0.000	0.567	0.000	-0.129	0.017	-0.129	0.017	-0.129	0.017	-0.155	0.022	0.241	0.000	0.321	0.000
C12	0.726	0.000	0.726	0.000	0.443	0.000	0.657	0.000	0.171	0.001	0.171	0.001	0.171	0.001	0.263	0.000	0.197	0.001	0.432	0.000
D1	0.822	0.000	0.822	0.000	0.502	0.000	0.470	0.000	0.332	0.000	0.332	0.000	0.332	0.000	0.323	0.000	0.252	0.000	0.221	0.000
D2	0.556	0.000	0.556	0.000	0.340	0.000	0.309	0.000	0.936	0.000	0.936	0.000	0.936	0.000	0.885	0.000	0.115	0.003	0.096	0.001
D3	0.522	0.000	0.522	0.000	0.319	0.000	0.290	0.000	1.190	0.000	1.190	0.000	1.190	0.000	1.125	0.000	0.101	0.005	0.084	0.003
D4	0.385	0.000	0.385	0.000	0.235*	0.000	0.230*	0.000	0.778	0.000	0.778	0.000	0.778	0.000	0.793	0.000	0.055	0.013	0.053	0.003
D5	0.556	0.000	0.556	0.000	0.339	0.000	0.444	0.000	0.695	0.000	0.695	0.000	0.695	0.000	0.946	0.000	0.115	0.003	0.197	0.000
D6	0.628	0.000	0.628	0.000	0.383	0.000	0.317	0.000	1.336	0.000	1.336	0.000	1.336	0.000	1.149	0.000	0.147	0.000	0.101	0.000
D7	0.347	0.000	0.347	0.000	0.212*	0.000	0.189*	0.000	0.734	0.000	0.734	0.000	0.734	0.000	0.680	0.000	0.045	0.007	0.036	0.010
D8	0.510	0.000	0.510	0.000	0.311	0.000	0.284	0.000	1.200	0.000	1.200	0.000	1.200	0.000	1.138	0.000	0.097	0.000	0.081	0.001
D9	0.378	0.000	0.378	0.000	0.231*	0.000	0.299	0.000	0.960	0.000	0.960	0.000	0.960	0.000	1.291	0.000	0.053	0.016	0.089	0.000
D10	0.269	0.000	0.269	0.000	0.164*	0.000	0.159*	0.000	1.400	0.000	1.400	0.000	1.400	0.000	1.411	0.000	0.027	0.041	0.025	0.036
D11	0.684	0.000	0.684	0.000	0.417	0.000	0.429	0.000	0.468	0.000	0.468	0.000	0.468	0.000	0.499	0.000	0.174	0.000	0.184	0.000
D12	0.679	0.000	0.679	0.000	0.414	0.000	0.373	0.000	1.013	0.000	1.013	0.000	1.013	0.000	0.949	0.000	0.172	0.000	0.139	0.000
E1	0.507	0.000	0.507	0.000	0.309	0.000	0.324	0.000	0.614	0.000	0.614	0.000	0.614	0.000	0.668	0.000	0.096	0.011	0.105	0.001
E2	0.785	0.000	0.785	0.000	0.479	0.000	0.503	0.000	1.055	0.000	1.055	0.000	1.055	0.000	1.152	0.000	0.229	0.000	0.253	0.000
E3	-0.001	0.993	-0.001	0.993	0.000*	0.993**	0.000*	0.993**	1.565	0.000	1.565	0.000	1.565	0.000	1.517	0.000	0.000	0.997	0.000	0.997
E4	0.007	0.922	0.007	0.922	0.004*	0.922**	0.004*	0.922**	1.113	0.000	1.113	0.000	1.113	0.000	1.048	0.000	0.000	0.961	0.000	0.961
E5	0.121	0.119	0.121	0.119	0.074*	0.109**	0.058*	0.111**	1.292	0.000	1.292	0.000	1.292	0.000	1.056	0.000	0.005	0.423	0.003	0.426
E6	-0.034	0.668	-0.034	0.668	-0.021*	0.669**	-0.019*	0.669**	1.440	0.000	1.440	0.000	1.440	0.000	1.375	0.000	0.000	0.831	0.000	0.831
E7	0.023	0.780	0.023	0.780	0.014*	0.780**	0.012*	0.780**	1.560	0.000	1.560	0.000	1.560	0.000	1.395	0.000	0.000	0.889	0.000	0.889
E8	0.207	0.008	0.207	0.008	0.126*	0.006	0.109*	0.009	1.469	0.000	1.469	0.000	1.469	0.000	1.321	0.000	0.016	0.168	0.012	0.189
E9	0.180	0.016	0.180	0.016	0.110*	0.013	0.115*	0.010	1.425	0.000	1.425	0.000	1.425	0.000	1.552	0.000	0.012	0.214	0.013	0.199
E10	-0.269	0.010	-0.269	0.010	-0.164*	0.006	-0.130*	0.012	1.689	0.000	1.689	0.000	1.689	0.000	1.389	0.000	0.027	0.170	0.017	0.209
E11	1.000	999.000	1.000	999.000	0.610	0.000	0.659	0.000	-1.842	0.000	-1.842	0.000	-1.842	0.000	-2.067	0.000	0.372	0.000	0.434	0.004
E12	1.032	0.000	1.032	0.000	0.630	0.000	0.568	0.000	-1.944	0.000	-1.944	0.000	-1.944	0.000	-1.822	0.000	0.396	0.000	0.322	0.011

*Item factor loading did not reach the minimum required threshold of 0.3 (McDonald, 1999).

**Item factor loading did not reach statistical significance.

Table 27: Unstandardised Parameter Estimates, Indicator Thresholds, and Proportion of Variance Explained (R^2) for Older Males and Older Females

	Male		Female		Male		Female		Male		Female		Male		Female		Male		Female	
Indicator	Unstand. Factor Loading	*p*	Unstand. Factor Loading	*p*	Standardised Factor Loading	*p*	Standardised Factor Loading	*p*	Unstand. Indicator Threshold	*p*	Unstand. Indicator Threshold	*p*	Standardised Indicator Threshold	*p*	Standardised Indicator Threshold	*p*	Prop. of Variance Explained (R^2)	*p*	Prop. of Variance Explained (R^2)	*p*
A3	1.000	0.000	1.000	0.000	0.745	0.000	0.778	0.000	-1.842	0.000	-1.842	0.000	-1.833	0.000	-2.135	0.000	0.555	0.000	0.605	0.024
A4	1.032	0.000	1.032	0.000	0.744	0.001	0.771	0.000	-1.944	0.000	-1.944	0.000	-1.873	0.000	-2.164	0.000	0.554	0.084	0.595	0.055
A5	0.932	0.000	0.932	0.000	0.631	0.000	0.799	0.000	-1.601	0.000	-1.601	0.000	-1.448	0.000	-2.043	0.000	0.398	0.000	0.638	0.060
A6	0.904	0.000	0.904	0.000	0.427	0.000	0.837	0.000	-1.282	0.000	-1.282	0.000	-0.808	0.000	-1.768	0.000	0.182	0.000	0.700	0.045
A7	1.137	0.000	1.137	0.000	0.544	0.000	0.569	0.000	-1.024	0.000	-1.024	0.000	-0.655	0.000	-0.762	0.000	0.296	0.000	0.323	0.000
A8	1.187	0.000	1.187	0.000	0.815	0.000	0.419	0.000	-1.528	0.000	-1.528	0.000	-1.401	0.000	-0.802	0.000	0.665	0.063	0.175	0.011
A9	1.164	0.000	1.164	0.000	0.616	0.000	0.467	0.000	-1.091	0.000	-1.091	0.000	-0.771	0.000	-0.652	0.000	0.379	0.000	0.218	0.011
A10	0.922	0.000	0.922	0.000	0.441	0.000	0.511	0.000	-0.595	0.000	-0.595	0.000	-0.381	0.000	-0.491	0.000	0.194	0.004	0.261	0.000
A11	0.530	0.000	0.530	0.000	0.400	0.000	0.490	0.000	-0.166	0.002	-0.166	0.002	-0.168	0.002	-0.229	0.001	0.160	0.014	0.240	0.000
A12	1.173	0.000	1.173	0.000	0.438	0.000	0.572	0.000	-1.897	0.000	-1.897	0.000	-0.947	0.000	-1.378	0.000	0.192	0.000	0.327	0.000
B1	0.947	0.000	0.947	0.000	0.564	0.000	0.809	0.000	-1.490	0.000	-1.490	0.000	-1.187	0.000	-1.897	0.000	0.318	0.008	0.655	0.002
B2	1.301	0.000	1.301	0.000	0.686	0.000	0.783	0.000	-1.014	0.000	-1.014	0.000	-0.715	0.000	-0.910	0.000	0.471	0.000	0.614	0.000
B3	1.466	0.000	1.466	0.000	0.864	0.000	0.759	0.000	-1.258	0.000	-1.258	0.000	-0.990	0.000	-0.970	0.000	0.746	0.020	0.576	0.000
B4	1.194	0.000	1.194	0.000	0.426	0.000	0.500	0.000	-0.687	0.000	-0.687	0.000	-0.328	0.000	-0.429	0.000	0.182	0.005	0.250	0.001
B5	0.925	0.000	0.925	0.000	0.400	0.000	0.330	0.000	-0.568	0.000	-0.568	0.000	-0.328	0.000	-0.302	0.000	0.160	0.006	0.109	0.024
B6	1.220	0.000	1.220	0.000	0.624	0.000	0.730	0.000	-0.749	0.000	-0.749	0.000	-0.511	0.000	-0.667	0.000	0.389	0.000	0.532	0.000
B7	1.149	0.000	1.149	0.000	0.441	0.000	0.452	0.000	-0.873	0.000	-0.873	0.000	-0.448	0.000	-0.512	0.000	0.195	0.001	0.204	0.011
B8	1.315	0.000	1.315	0.000	0.684	0.000	0.575	0.000	-1.022	0.000	-1.022	0.000	-0.711	0.000	-0.665	0.000	0.468	0.000	0.330	0.000
B9	1.060	0.000	1.060	0.000	0.365	0.000	0.407	0.000	-0.770	0.000	-0.770	0.000	-0.354	0.000	-0.440	0.000	0.133	0.011	0.166	0.001
B10	0.957	0.000	0.957	0.000	0.424	0.000	0.441	0.000	-0.273	0.000	-0.273	0.000	-0.162	0.000	-0.188	0.000	0.180	0.004	0.194	0.002
B11	1.296	0.000	1.296	0.000	0.703	0.000	0.659	0.000	-1.184	0.000	-1.184	0.000	-0.858	0.000	-0.896	0.000	0.494	0.000	0.434	0.031
B12	0.948	0.000	0.948	0.000	0.557	0.000	0.592	0.000	-1.135	0.000	-1.135	0.000	-0.891	0.000	-1.056	0.000	0.310	0.000	0.350	0.000
C1	0.969	0.000	0.969	0.000	0.398	0.000	0.555	0.000	-0.507	0.000	-0.507	0.000	-0.279	0.000	-0.433	0.000	0.159	0.004	0.308	0.000
C2	1.144	0.000	1.144	0.000	0.686	0.000	0.815	0.000	-1.308	0.000	-1.308	0.000	-1.047	0.000	-1.387	0.000	0.470	0.000	0.664	0.000
C3	0.999	0.000	0.999	0.000	0.709	0.000	0.577	0.000	-0.051	0.413	-0.051	0.413	-0.048	0.411	-0.044	0.411	0.503	0.000	0.332	0.000
C4	1.034	0.000	1.034	0.000	0.535	0.000	0.455	0.000	-0.753	0.000	-0.753	0.000	-0.520	0.000	-0.493	0.000	0.286	0.001	0.207	0.000
C5	0.636	0.000	0.636	0.000	0.523	0.000	0.373	0.000	0.497	0.000	0.497	0.000	0.546	0.000	0.434	0.000	0.274	0.002	0.139	0.042
C6	0.916	0.000	0.916	0.000	0.528	0.000	0.616	0.000	0.064	0.277	0.064	0.277	0.049	0.283	0.064	0.285	0.278	0.000	0.380	0.000
C7	0.646	0.000	0.646	0.000	0.395	0.000	0.402	0.000	0.970	0.000	0.970	0.000	0.792	0.000	0.900	0.000	0.156	0.036	0.162	0.019
C8	0.167	0.021	0.167	0.021	0.147*	0.033	0.131*	0.026	1.222	0.000	1.222	0.000	1.437	0.000	1.425	0.000	0.022	0.286	0.017	0.266

*Item factor loading did not reach the minimum required threshold of 0.3 (McDonald, 1999).

**Item factor loading did not reach statistical significance.

	Male		Female		Male		Female		Male		Female		Male		Female		Male		Female	
Indicator	Unstand. Factor Loading	*p*	Unstand. Factor Loading	*p*	Standardised Factor Loading	*p*	Standardised Factor Loading	*p*	Unstand. Indicator Threshold	*p*	Unstand. Indicator Threshold	*p*	Standardised Indicator Threshold	*p*	Standardised Indicator Threshold	*p*	Prop. of Variance Explained (R^2)	*p*	Prop. of Variance Explained (R^2)	*p*
C9	0.484	0.000	0.484	0.000	0.386	0.000	0.469	0.000	1.319	0.000	1.319	0.000	1.405	0.000	1.903	0.000	0.149	0.010	0.220	0.001
C10	0.653	0.000	0.653	0.000	0.310	0.001	0.253	0.001	1.301	0.000	1.301	0.000	0.824	0.000	0.750	0.000	0.096	0.101	0.064	0.102
C11	0.805	0.000	0.805	0.000	0.481	0.000	0.517	0.000	-0.129	0.017	-0.129	0.017	-0.103	0.019	-0.123	0.018	0.232	0.001	0.268	0.000
C12	0.726	0.000	0.726	0.000	0.700	0.000	0.722	0.000	0.171	0.001	0.171	0.001	0.220	0.001	0.253	0.001	0.490	0.000	0.521	0.000
D1	0.822	0.000	0.822	0.000	0.649	0.000	0.632	0.000	0.332	0.000	0.332	0.000	0.350	0.000	0.380	0.000	0.421	0.000	0.399	0.000
D2	0.556	0.000	0.556	0.000	0.477	0.000	0.278	0.001	0.936	0.000	0.936	0.000	1.071	0.000	0.695	0.000	0.227	0.005	0.077	0.093
D3	0.522	0.000	0.522	0.000	0.331	0.000	0.324	0.000	1.190	0.000	1.190	0.000	1.007	0.000	1.098	0.000	0.109	0.057	0.105	0.013
D4	0.385	0.000	0.385	0.000	0.431	0.000	0.351	0.000	0.778	0.000	0.778	0.000	1.164	0.000	1.055	0.000	0.186	0.023	0.123	0.021
D5	0.556	0.000	0.556	0.000	0.695	0.000	0.550	0.000	0.695	0.000	0.695	0.000	1.161	0.000	1.025	0.000	0.483	0.000	0.303	0.000
D6	0.628	0.000	0.628	0.000	0.526	0.000	0.191*	0.004	1.336	0.000	1.336	0.000	1.496	0.000	0.605	0.002	0.277	0.001	0.036	0.153
D7	0.347	0.000	0.347	0.000	0.513	0.000	0.263	0.001	0.734	0.000	0.734	0.000	1.447	0.000	0.828	0.000	0.263	0.001	0.069	0.092
D8	0.510	0.000	0.510	0.000	0.515	0.000	0.288	0.000	1.200	0.000	1.200	0.000	1.617	0.000	1.008	0.000	0.265	0.000	0.083	0.040
D9	0.378	0.000	0.378	0.000	0.493	0.000	0.330	0.000	0.960	0.000	0.960	0.000	1.671	0.000	1.246	0.000	0.243	0.002	0.109	0.005
D10	0.269	0.000	0.269	0.000	0.165*	0.004	0.165*	0.000	1.400	0.000	1.400	0.000	1.149	0.000	1.283	0.000	0.027	0.155	0.027	0.067
D11	0.684	0.000	0.684	0.000	0.558	0.000	0.452	0.000	0.468	0.000	0.468	0.000	0.510	0.000	0.460	0.000	0.312	0.001	0.204	0.008
D12	0.679	0.000	0.679	0.000	0.652	0.000	0.462	0.000	1.013	0.000	1.013	0.000	1.300	0.000	1.027	0.000	0.425	0.000	0.213	0.001
E1	0.507	0.000	0.507	0.000	0.512	0.000	0.580	0.000	0.614	0.000	0.614	0.000	0.829	0.000	1.046	0.000	0.262	0.004	0.336	0.000
E2	0.785	0.000	0.785	0.000	0.693	0.000	0.648	0.000	1.055	0.000	1.055	0.000	1.245	0.000	1.297	0.000	0.481	0.000	0.420	0.000
E3	-0.001	0.993	-0.001	0.993	0.000*	0.993**	0.000*	0.993**	1.565	0.000	1.565	0.000	1.118	0.000	1.497	0.000	0.000	0.997	0.000	0.997
E4	0.007	0.922	0.007	0.922	0.005*	0.922**	0.005*	0.922**	1.113	0.000	1.113	0.000	1.005	0.000	1.086	0.000	0.000	0.961	0.000	0.961
E5	0.121	0.119	0.121	0.119	0.093*	0.132**	0.082*	0.133**	1.292	0.000	1.292	0.000	1.323	0.000	1.300	0.000	0.009	0.452	0.007	0.452
E6	-0.034	0.668	-0.034	0.668	-0.026*	0.665**	-0.026*	0.669**	1.440	0.000	1.440	0.000	1.454	0.000	1.625	0.000	0.001	0.829	0.001	0.831
E7	0.023	0.780	0.023	0.780	0.019*	0.779**	0.012*	0.782**	1.560	0.000	1.560	0.000	1.702	0.000	1.167	0.000	0.000	0.889	0.000	0.890
E8	0.207	0.008	0.207	0.008	0.153*	0.012	0.150*	0.009	1.469	0.000	1.469	0.000	1.452	0.000	1.587	0.000	0.023	0.208	0.022	0.191
E9	0.180	0.016	0.180	0.016	0.111*	0.021	0.142*	0.013	1.425	0.000	1.425	0.000	1.174	0.000	1.673	0.000	0.012	0.247	0.020	0.214
E10	-0.269	0.010	-0.269	0.010	-0.145*	0.009	-0.152*	0.005	1.689	0.000	1.689	0.000	1.215	0.000	1.420	0.000	0.021	0.191	0.023	0.157
E11	1.000	999.000	1.000	999.000	0.745	0.000	0.778	0.000	-1.842	0.000	-1.842	0.000	-1.833	0.000	-2.135	0.000	0.555	0.000	0.605	0.024
E12	1.032	0.000	1.032	0.000	0.744	0.001	0.771	0.000	-1.944	0.000	-1.944	0.000	-1.873	0.000	-2.164	0.000	0.554	0.084	0.595	0.055

*Item factor loading did not reach the minimum required threshold of 0.3 (McDonald, 1999).

**Item factor loading did not reach statistical significance.

Examination of the standardised factor loadings reveals that 51 indicator factor loadings were statistically significant for younger males, younger females, older males, and older females: p < 0.01. The latent factor explained significant variance: $R^2_{Younger\ males}$ mean = 0.259, range = 0 to 0.800; $R^2_{Younger\ females}$ mean = 0.246, range = 0 to 0.763; $R^2_{Older\ males}$ mean = 0.264, range = 0 to 0.746; $R^2_{Older\ females}$ mean = 0.256, range = 0 to 0.700. A number of the items did not reach statistical significance or salience with the suggested minimum threshold of 0.3 (McDonald, 1999). Indicators that fail to significantly and saliently load are not considered to be well accounted for by the latent factor. The non-significant and non-salient items will now be detailed for each of the four groups separately.

For younger males (Table 26) 8 of the 56 test items did not reach salience, but reached significance. A further 8 items did not reach either salience or significance. For younger females (Table 26), 8 of the 56 test items did not reach salience, but reached significance. A further 8 items did not reach either salience or significance. For the older males (Table 27), 5 of the 56 test items reached significance but not salience, while 5 items were not salient or significant. Finally, for the older females (Table 27), 5 of the 56 test items reached significance but not salience. An additional 6 items reached neither salience nor significance.

To determine whether the elimination of the non-significant and non-salient items from the MG-CFA would significantly improve the one-factor model fit for four groups, the 39 significant and salient items were loaded onto the one latent factor. The model (M7-1a) was over-identified with 205 *df*; χ^2 (205) = 325.394, p =0.001, CFI = 0.908, TLI = 0.916, RMSEA = 0.050. A chi-square difference test was performed: $\chi^2_{\text{diff}}(92)$ = 160.652 (the critical value of chi-square at 92 *df* is 115.39, p = 0.05). From this it can be concluded that the four-group model would be significantly improved by the removal of the non-significant and non-salient items from model M7-1.

Inspection of the remaining 39 test items revealed that items C5, and C8 did not have significant indicator thresholds. An indicator threshold provides the value of the indicator when the latent factor is zero, and when this value is significant, is indication that the value is significantly different than zero. Items C5 and C8 cannot be viewed as being significantly different to zero. However, in the present context, items C5 and C8 being non-significantly different to zero can be interpreted as being the approximate midpoint in the range of threshold values.

For each of the four groups, 27 of the 39 test thresholds were negative and 12 of the 39 were positive. The indicator thresholds can be interpreted in relation to item difficulty that was discussed in Chapters 5 and 6. The difficulty of each indicator is represented along a continuum of thresholds, ranging from the easiest item for participants (item A4 for older females) with a threshold of -2.164, to the most difficult item (item C11 for older females), with a threshold of 1.903.

Upon further examination of the non-suitable items of the one-factor solution, many of the non-significant and non-salient items are common across the four groups. These findings suggest that the non-significant and/or non-salient items may not be sufficiently represented by the 1-factor model of general intelligence. This may be the result of imposing equality constraints upon the model, or it may be that the differential covariance amongst the items may be attributable to the measurement approach. This will be investigated by the addition of methods factors to the one-factor model in a later section of this chapter.

Despite of the low communalities and non-significance of some items in the model, for reasons of comparability and generalisability discussed previously analysis of mean and variance differences of a one-factor model with four groups will proceed using all items of the SPM+.

7.3. Assessment of Sex Differences in Latent Mean and Variability in Younger & Older Participants

The previously established one-factor MG-CFA (model M7-1) will now be expanded to assess sex differences in latent means and variability. This will be first described for younger and older males, and for younger and older females. These analyses will be provided as a verification of validity in order to better understand the performance of younger versus older participants. Comparisons of mean and variance will then be discussed for younger males and females followed by a discussion for older males and females.

First, the variability of the latent factor will be constrained to equality for the younger and older males (model M7-2) to assess differences in the latent factor mean. Next, the latent factor mean of younger males will be fixed to zero in order to identify the mean structure component of the MG-CFA. In doing so, the latent factor mean of younger males will serve as a reference, while the latent factor mean of the older males will represent the difference between the two latent factor means. The same analyses will be repeated for younger and older females, for younger males and females, and finally, for older males and females.

7.3.1. Younger and Older Males and Females

In order to examine the sex differences in the latent mean, the variability of the latent factor was held to equality for younger and older males, while the variability for younger and older females was allowed to be freely estimated. The model (M7-2) was over-identified with 296 *df*; χ^2 (296) = 484.880, $p < 0.001$. The fit statistics indicate a good model fit: CFI = 0.839, TLI = 0.846, RMSEA = 0.054. The latent mean was 1.373 standard deviations higher for older males than for younger males. This difference was significant: z = 10.586, $p < 0.001$.

In order to examine the sex differences in the latent mean for younger and older females, the variability of the latent factor was held to equality, while the variability for younger and older males was allowed to be freely estimated. The model (M7-3) was over-identified with 296 *df*; χ^2 (296) = 486.074, $p < 0.001$. The fit statistics indicate a good model fit: CFI = 0.839, TLI = 0.847, RMSEA = 0.052. The latent mean was 1.387 standard deviations higher for older females than for younger females. This difference was significant: z = 11.252, p < 0.001.

In relation to variance, the previous, measurement invariance model (M7-1) solution showed that the overall variance was 0.373 for younger boys and 0.556 for older males. The fit of model M7-2 was compared to that of the previous, measurement invariance model, M7-1. This analysis tests whether the SPM+ items drew on similar ranges of the latent factor of general intelligence for younger and older males. There was not a significant degradation to the model fit when the latent factor mean of general intelligence was held to equality for younger and older males: χ^2_{diff} (1) = 1.166 (the critical value of chi-square at 1 *df* is 3.84, *p* = 0.05). From this, it can be concluded that the variance in general intelligence, as measured by the SPM+, is not significantly different between younger and older males.

For younger and older females, the previous, measurement invariance model (M7-1) solution showed that the overall variance was 0.350 for younger females and 0.458 for older females. The fit of model M7-3 was compared to the previous, measurement invariance model, M7-1. As with the males, there was not a significant degradation to the model fit when the latent factor mean of general intelligence was held to equality for younger and older females: χ^2_{diff} (1) = 0.028 (the critical value of chi-square at 1 *df* is 3.84, *p* = 0.05). From this, it can be concluded that the variance in general intelligence, as measured by the SPM+, is not significantly different between younger and older females.

To summarise, the means of the older participants are significantly greater than for the younger participants. However, the variance is found to be equivalent when comparing younger to older males and females.

7.3.2. Younger Males and Females

The variability of the latent factor of model M7-1 was held to equality for younger males and females, while the variability for older males and females was allowed to be freely estimated, in order to examine the sex differences in the latent mean. The model (M7-4) was over-identified with 297 *df*; χ^2 (297) = 485.173, $p < 0.001$. The fit statistics indicate a good model fit: CFI = 0.839, TLI = 0.847, RMSEA = 0.052. The latent mean was 0.01 standard deviations higher for younger girls than for younger boys. This difference was not significant: z = 0.086, p = 0.931.

With respect to variance, the previous, measurement invariance model (M7-1) solution showed that the overall variance was 0.373 for younger boys and 0.350 for younger girls. The fit of model M7-4 with constrained factor loadings was compared to that of the previous, measurement invariance model, M7-1. This comparison verifies whether the SPM+ items drew on similar ranges of the latent factor of general intelligence for younger boys and girls.

The model fit was not significantly degraded when the latent factor of general intelligence was held to equality for younger boys and girls: χ^2_{diff} (1) = 0.474 (the critical value of chi-square at 1 *df* is 3.84, $p = 0.05$). From this, it can be concluded that the variance in general intelligence, as measured by the SPM+, is not significantly different between younger boys and girls.

7.3.3. Older Males and Females

The variability of the latent factor of model M7-1 was held to equality for older males and females, while the variability for younger males and females was allowed to be freely estimated. The model (M7-5) was over-identified with 297 *df*; χ^2 (297) = 485.989, $p < 0.001$. The fit statistics indicate a good model fit: CFI = 0.839, TLI = 0.847, RMSEA = 0.052. The latent mean was .026 standard deviations higher for older girls than for older boys. This difference was not significant: z = 0.194, p = 0.846.

With respect to variance, the measurement invariance model (M7-1) solution showed that the overall variance was 0.556 for older boys and 0.458 for older girls. Again, the fit of the extended model (M7-5) was compared to that of the previous, measurement invariance model, M7-1. The model fit was not significantly degraded when the latent factor of general intelligence was held to equality for older boys and girls: χ^2_{diff} (1) = 0.741 (the critical value of chi-square at 1 *df* is 3.84, $p = 0.05$). It can be concluded that the variance in general intelligence, as measured by the SPM+, is not significantly different between older boys and girls. From this, it can be interpreted that the two groups are drawing from the same range of ability when answering items on the SPM+.

7.4. Methods Factor

In the models presented thus far, a certain amount of variance has been unexplained by the one-factor solution. Across the large age range of the standardisation sample of the SPM+, it is possible that performance could be influenced by extraneous elements, such as a methods factor. For this reason, it is reasonable to further expand upon model M7-1 to evaluate whether there is an effect of methods factors between males and females in the younger and older age groups.

When the model was specified for the younger males and females (58 test items loaded onto one latent factor and 18 items loaded onto two additional methods factors), the model would not converge. Model non-convergence can occur for a number of reasons, the most common of which is when the specified model is not supported by the data. This is very often related to the model complexity, with large numbers of freely estimated parameters (Brown, 2006). This seems to be the case with the present model as it is not possible to estimate a MG-CFA with two methods factors for younger males and females for the U.K. standardisation data of the SPM+.

However, specification of this complex model was possible for the older participants. 56 SPM+ test items were loaded onto one latent factor, with an additional 18 items loaded onto two secondary methods factors. Upon inspection of the modification indices, it was suggested to remove five of the 18 items on the 'Gestalt' method factor. These items were found to be 'non-positive definite', where there is negative variance/residual variance (Brown, 2006), and were subsequently removed from the analyses.

The model (M7-6) was specified for older males and females, with 56 items on one latent factor and 13 items on two methods factors (Gestalt and Visuospatial), conceptualised in accordance with the existing literature on item solving strategies (as discussed in Chapter 6). The model was over-identified with 102 *df*; χ^2 (102) = 164.032, $p < 0.001$. The statistics indicated adequate model fit: CFI = 0.716, TLI = 0.719, RMSEA = 0.068. No further points of ill-fit were identified in the model.

In order to compare the means of the methods factors for males and females, the variability of the factors is held to equality. The model (M7-7) was over-identified with 102 *df*; χ^2 (102) = 164.229, $p < 0.001$. The statistics indicated adequate model fit: CFI = 0.715, TLI = 0.718, RMSEA = 0.068. The method factor mean for Gestalt items was 0.138 standard deviations higher for males than for females, but this difference was not significant: z = -0.210, *p* = 0.834. The methods factor mean for Visuospatial items was 0.703 standard deviations higher for females than for males, but again, this difference was not significant: z = 1.411, *p* = 0.158.

Table 28. Summary of Model Fit Statistics and Chi-square Difference Tests

Model	**Chi-square statistic (χ^2)**	**Degrees of Freedom (*df*)**	**CFI**	**RMSEA**	**Chi-square Difference (χ^2_{diff})**	**Degrees of Freedom Difference (*df*$_{diff}$)**	***p***
M7-1	486.046	297	0.839	0.052			
M7-1a	325.394	205	0.908	0.050	160.652[24]	92	SIG
M7-2	484.880	296	0.839	0.054	1.166[25]	1	NS
M7-3	486.074	296	0.839	0.052	0.028[26]	1	NS
M7-4	485.173	297	0.839	0.052	.741[27]	1	NS
M7-5	485.989	297	0.839	0.052	.474[28]	1	NS
M7-6	164.032	102	0.716	0.068			
M7-7	164.229	102	0.715	0.068			

7.5. Summary

The issue of sex differences in general intelligence (as measured by the SPM+) has been further assessed by accounting for the varied age of the sample participants. In doing so, the third and fourth objectives of

[24] The critical value of chi-square at 92 *df* is 115.39, *p*=0.05.

[25] The critical value of the chi-square at 1 *df* is 3.84, p=0.05.

[26] The critical value of the chi-square at 1 *df* is 3.84, p=0.05.

[27] The critical value of the chi-square at 1 *df* is 3.84, p=0.05.

[28] The critical value of the chi-square at 1 *df* is 3.84, p=0.05.

this dissertation were addressed. A one-factor model of a latent factor of general intelligence was assessed for four groups of participants from the U.K. standardisation sample: younger males, younger females, older males, and older females. As evidenced from the goodness-of-fit statistics, the U.K. SPM+ standardisation data can be represented sufficiently well by a one-factor solution.

Evidence from the least constrained, measurement invariance model (M7-1) indicates that there are no significant differences in factor loadings between the four groups. In combination with the previously established equivalence of indicator thresholds, it can be interpreted that the metric for the four groups of participants at each value of the latent factor is equivalent. In other words, items on the SPM+ are measuring general intelligence in the same way for boys and girls across the younger and older age groups.

Having confirmed this, it was appropriate to proceed with an assessment of sex differences in the latent factor mean and variance in each age group. Variance of the latent factors was constrained to equality for males and females in each age group. First a comparison was made between younger and older participants. As expected, the older participants achieved a significantly greater mean score that the younger participants.

When comparing the sexes in the group of younger participants, girls achieved 0.01 standard deviations higher than boys, but not significantly so. When comparing the sexes in the group of older participants, girls achieved 0.026 standard deviations higher than boys. Again this was not a significant difference.

When equality constraints were imposed upon the model to assess group differences in variance, there was no significant degradation to the model for either age group. This indicates that the variance in the latent factor is not significantly different between younger boys and girls, or between older boys and girls.

One unexpected finding from the one-factor solution is that there were a number of items that did not have significant and/or salient factor loadings. Upon further inspection, many items are common across all four of the groups in the model. When these items were removed from the model, there was a significant improvement in global model fit. This finding may suggest that some of the covariance among these items is related to something other than the latent factor of general intelligence, such as the measurement approach or the item type. To assess this, the one factor model was extended to include two methods factors.

The one-factor MG-CFA methods factor model would not converge for the younger group of males and females. For the older participants the measurement invariance model indicates that the metric of the methods factors is equivalent for boys and girls. In other words, the influence of the measurement approach is the same for older boys and girls. When each of the methods factors was isolated in succession, it was revealed that there was no significant difference in the mean of the first methods factor (or what might be thought of as Gestalt items) or in the mean of the second methods factor (or what might be thought of as Visuospatial items). These results must be interpreted with caution due to the lack of power to reject the null hypothesis resulting from less than optimal sample size of the group of older participants who completed the SPM+.

To this point, the SPM+ has been assessed using a one-factor model in a number of different ways. In Chapter 5, a one-factor solution was assessed using a Rasch model, illustrating that while the difficulty of a number of test items do not follow the order of presentation, the average item set difficulty is the same for males and females. In chapter 6, a one-factor CFA model was used to illustrate that the SPM+ measures the construct of general intelligence in the same way for males and females, and that there are no mean differences between the two sexes. These findings were extended further in this chapter. Four-group MG-CFA analyses revealed that, for both the younger and older participants of the sample, there remains no differences in mean or variability in general intelligence as measured by the SPM+.

— 8 —

Discussion

8.1. Introduction

The findings from this dissertation make novel contributions to the extensive literature by employing robust statistical modelling techniques to explore sex differences in general intelligence using a large representative sample of the U.K. standardisation of the Raven's Standard Progressive Matrices. In addressing the four main research objectives, this dissertation provides some balance to what has been largely a one-sided debate arguing that men have superior general intelligence to women (Begley, 2009).

In accordance with the objectives of this study, four main findings emerged from this assessment of sex differences in general intelligence as measured by the Raven's SPM+. First, it was determined that the measurement properties for each value of the latent factor are equivalent for boys and girls in the overall sample as well as in the younger and older age groups. That is to say that the items on the SPM+ are measuring general intelligence in the same way for both boys and girls in the overall sample as well as at younger and older ages.

Second, the mean scores on the SPM+ were not found to be significantly different between boys and girls. Further, the scores were found to be equally variable in both males and females in the sample as a whole.

Third, the mean scores on the SPM+ were not found to be significantly different between boys and girls in the younger or older age groups of participants. As with the overall sample, the males and females were found to be equally variable at younger and older ages.

Finally, it was determined that once the latent factor of general intelligence was accounted for in the overall sample, males were slightly more negatively affected by the visuospatial element of some of the easier items on the measure. However, this effect was no longer apparent when the effects of the methods factors were assessed by age group.

This chapter will provide a discussion of each of these findings in turn (sections 8.3, 8.4, and 8.5). It concludes with a discussion of the strengths and limitations of the study's findings in relation to the existing literature, and the implications for further investigation (section 8.6 and 8.7).

8.2. Male & Female Measurement Equivalence and Sampling issues in the SPM+

The first aim of this dissertation was to determine whether the Raven's SPM+ was biased towards either males or females; that is, whether the measurement properties of each value of the latent factor are equivalent for boys and girls. A secondary objective was to assess measurement equivalence in both the overall sample as well as in the younger and older age groups of participants.

Studies assessing sex differences on the SPM have, to date, generally made use of classical statistical methods to analyse the results, such as t-tests (Abdel-Khalek & Lynn, 2006; Arden & Plomin, 2006; Crucian & Berenbaum, 1998; Mohan & Kumar, 1979) and analysis of variance (Rushton & Skuy, 2000; Silverman et al., 2000). Such methods have been found to lack the required strength and sensitivity, and are no longer considered sufficient to effectively and accurately identify group differences (Embretson & Reise, 2000). As detailed in Chapter 4, classical methodologies summarise the properties of a test by a single omnibus statistic, such as Cronbach's α which is based upon correlations between different items on a test. In contrast, modelling techniques account for the variation in endorsing an item as a function of the individual's level of the latent construct in relation to the item parameters or characteristics (Baker, 2001; Santor et al., 1994). Further, many classical methods used in this literature (such as correlation analyses and analysis of variance) make the assumption that the data is free from error, which is rarely the case in social sciences. Modelling methods account for measurement error such that the resulting relationships between variables and the latent trait can be estimated after error has been adjusted.

In order to address the question of group differences responsibly and appropriately, the investigation must begin with an assessment of measurement equivalence to ensure that the measurement metric is the same for both groups. Without ensuring measurement invariance of a measure, it is unclear whether mean differences between groups are a genuine reflection of differences in the underlying trait, or if these differences are attributable to the bias within the measure (Embretson & Reise, 2000; Horn & McArdle, 1992; Keith et al., 2008).

As discussed in Chapter 4, if the same score on a test of intelligence is not representative of the same level of ability in different groups, a test is considered to be biased (Drasgow, 1984, Horn & McArdle, 1992). This is also referred to as Differential Item Functioning (DIF). DIF refers to instances where an item on a test yields a different mean response for members of different groups with the same latent trait score (e.g. intelligence).

The determination of measurement invariance ensures that any differences found between groups are in fact genuine differences in the latent construct rather than artefacts arising from measurement error or bias (van Der Sluis et al., 2008; Wicherts, Dolan, & Hessen, 2005). Establishment of measurement invariance is necessary to ensure that accurate conclusions about group differences are drawn (Horn & McArdle, 1992).

In addition to weak statistical methodology reported in prior studies, there is a further failure to ensure measurement equivalence and the possibility of item bias prior to the assessment of differences in mean and variability. It is apparent from the literature of sex differences on the SPM that these important preliminary verifications were not undertaken before proceeding with their assessment of mean differences.

It is therefore unclear from the sex differences noted in the SPM literature (Abdel-Khalek & Lynn, 2006; Arden & Plomin, 2006; Crucian & Berenbaum, 1998; Irwing & Lynn, 2005; Lynn & Irwing, 2004; Mohan &

Kumar, 1979; Rushton & Skuy, 2000; Silverman et al., 2000) whether the differences they found are attributable to actual differences in ability or whether the measure was biased against one of the groups in the populations that they assessed. The same could be said of the studies that failed to find a significant group difference (Rushton and Cvorovic, 2009; Lynn, Backhoff, & Contreras-Niño, 2004): if differential item functioning resulted in bias against one of the groups, then such studies failed, on psychometric grounds to find a true difference in ability.

For these reasons, the current study began with a thorough assessment of measurement properties of the SPM+ in order to ascertain whether the measure was fair to both boys and girls. This is thought to be a novel contribution to the literature. The MG-CFA techniques employed in this dissertation allowed for group comparisons of measurement characteristics (i.e., factor loadings, indicator intercepts or thresholds, residual variances) and structural features (i.e., factor variances, factor covariances, and latent means) of the latent factor which is of utmost importance to ensure that the measurement properties are equivalent in each group (Brown, 2006; van Der Sluis et al., 2008).

When the factor loadings and indicator thresholds were constrained to equality in the model, it was determined that the measurement properties of the indicators were equivalent for males and females. In other words, the SPM measured the latent construct of general intelligence the same way for males and females and was not biased against either group. This was true of the standardisation sample as a whole (section 6.2) as well as in younger and older sub-groups of the population (section 7.2).

In addition to a lack of robust statistical analyses in the literature, the sampling procedures to date have been less than optimal, with non-representative (or opportunity) samples having generally been used. Studies, such as Abdel-Khalek and Lynn (2006), and Lynn, Allik, Pullmann, et al., (2004), made use of samples that were readily available and not particularly representative of the general population. Further, in each of these studies the sample sizes were particularly large (N = 6529 and N = 2738, respectively). It has been noted that numerically large samples have the power to detect very small differences as significant. In conjunction with their use of classic statistical methodology, the issue remains whether the conclusion of significant differences are meaningful in reality. Their conclusions should, therefore, be interpreted with caution.

Extant findings could be distorted by biases introduced through sampling procedures (Molenaar, Dolan, & Wicherts, 2009). For example, in addition to using an opportunity sample, Rushton and Skuy (2000) did not use equal numbers of males and females in their samples for comparison (N_{female} = 205, N_{male} = 104). This was also found to be the case in other studies making claims of sex differences (Arden & Plomin, 2006; Rushton & Cvorovic, 2009; Silverman et al., 2000). Such unbalanced proportions of males and females may have introduced further sampling artefacts influencing the overall conclusions drawn by these studies.

In light of these sub-optimal analytic and sampling procedures, it is not considered appropriate to make generalisations to all men and women according to these study results. It is only when sound, representative research makes consistent findings are generalisation to more global populations appropriate (Halpern, In Press). In contrast, the use of modern statistical methodology with a large representative sample in the current dissertation allowed for the assessment of measurement properties in both male and female sub-groups of the sample as a whole as well as in younger and older age groups. This allowed for sound conclusions to be drawn of measurement invariance and the lack of bias in the U.K. Standardisation of the SPM+. As such, it is considered more appropriate to generalise the current findings to a larger population than from previous studies in the literature.

8.3. Sex Differences in Mean and Variability of the Overall Sample

The second objective of this dissertation was to assess whether significant sex differences existed in mean and variability of scores on the SPM+. As the Raven's Matrices are considered by many to be one of the best measures of general intelligence, this research objective can also be interpreted as assessing sex differences in mean and variability in general intelligence. The MG-CFA analyses conducted in this dissertation revealed that girls obtain 0.2 standard deviations higher on the SPM+ than boys, but this difference was not significant. The failure to find significant sex differences is concurrent with some previous findings, while contradictory to others.

The debate over "which is the smarter sex?" has continued, quite acrimoniously, for decades (Halpern, 2007). Since before the time when intelligence tests were first developed, females have been viewed in society as the weaker and the feebler of the sexes (Nature, 1923). As evidenced by a review of the current literature on the subject the perspective of females as 'inferior' to males has changed very little for some theorists.

An overwhelming majority of the studies of sex differences on the SPM report a male advantage (Arden & Plomin, 2006; Crucian & Berenbaum, 1998; Mohan & Kumar, 1979; Rushton & Skuy, 2000; Silverman et al., 2000), and are in direct opposition with the findings of the current analyses. Such studies have a tendency to maintain a "females-have-less" perspective (Halpern, In Press) in describing their results, interpreting their findings in such a way that implies females are inferior to males in terms of cognitive ability.

It could be argued at a number of levels that these studies are not providing the most statistically or methodologically sound assessments of the issue in question. The failure to assess potential sources of bias was described in section 2.4. The chosen statistical methods are also implicated. Each of these studies employed classical statistical methods, using t-tests and analysis of variance as proof of significant difference. These methods are now understood to lack the required strength and specificity for accurate detection of group differences. Further, the opportunity samples used in these studies are not deemed representative of the population at large. It is thus unsuitable to generalise from these findings to make claims about general intelligence of all males and females.

Meta-analyses of such studies have also arrived at a conclusion of male advantage (Irwing & Lynn, 2005; Lynn & Irwing, 2004). However, as detailed in section 2.3.2.2, the chosen methodology of these meta-analytic studies has been raised into question (Blinkhorn, 2005) due to the exclusion of large, representative samples, and the failure to uphold standard statistical practices (such as the failure to weight differences in score by sample size). Unfortunately, one of the largest and most up-to-date meta-analyses of the Progressive Matrices (Brouwers et al., 2009) was unable to assess the effect of sex, as a number of studies included in their analyses failed to report participant sex. Further, of the 798 samples included in the analyses, 175 were composed entirely of males and 113 composed entirely of females.

A rarity in the literature, Abdel-Khalek and Lynn (2006) identified a female advantage in a population of 8 to 15 year olds in Kuwait. Like the finding of male advantage, this finding is contradictory to the conclusions of the current investigation. Additionally, Abdel-Khalek and Lynn's finding is in accordance with the portion of Lynn's "Developmental Theory of Sex Differences" (1999) as it relates to younger children. This will be discussed further in section 8.4 in relation to the third research objective of this dissertation. While such results could be generalised to other 8 to 15 year old females in Kuwait, such a study provides little in the way of generalisability to other populations due to the way the sample was collected. A large opportunity sample such as this will undoubtedly be influenced by sociological factors such as parent's educational level and income that should be accounted for in the analyses and ultimately, in their overall conclusions about intelligence.

A further rarity in the literature concerning sex differences on the SPM is the conclusion that males and females perform equally well. In the current dissertation, MG-CFA analyses revealed that girls obtain scores 0.02 standard deviations higher on the SPM+ than boys, but this difference was not significant. Such a finding is corroborated by Rushton and Cvorovic (2009) and by Lynn, Backhoff, and Contreras-Niño (2004). However, the current finding is considered more robust than these prior studies due to the strength of the statistical methods and the quality of the representative sample employed.

Within the context of the larger literature on general intelligence as measured by other tests of intelligence, such as the WISC or the Naglieri Non-Verbal Ability Test, a lack of significant differences between males' and females' performance is concluded much more commonly (Camarata &

Woodcock, 2006; Colom, Juan-Espinosa, Abad, & García, 2000; van Der Sluis et al., 2008; Jensen, 1998; van der Sluis et al., 2006). Unlike the SPM, the measures used in these studies are generally comprised of a number of subtests each testing different cognitive abilities, such as processing speed or working memory. From a factor analytic perspective, such measures would conceive intelligence in terms of multiple factors of different intellectual abilities with a higher-order factor that is representative of general intelligence. It is important to note that a number of these studies made use of standardisation samples that were representative of the population. In those studies that did not use representative samples, the sampling strategies were such that equal numbers of males and females were used, and demographic variables (such as attained education levels) were accounted for. Attention to sampling quality may have contributed to the lack of findings of sex differences in cognitive ability.

While there has been considerable debate and great contention about the issue of mean differences in general intelligence, the debate about sex differences in variability has been slightly less discordant. A further finding from this dissertation was that, for the sample as a whole, scores on the SPM+ are equally variable in both males and females. This is contrary to an existing finding that determined females to be more variable on the SPM (Irwing & Lynn, 2005).

There is a relative lack of studies detailing the variance on the SPM, as it is only in the last 15 years that variability of scores has been considered in analyses (Hedges & Nowell, 1995; Nowell & Hedges, 1998), which is relatively recent in the intelligence literature that spans many decades. To date, investigations of differences in mean have been the priority, followed by the investigations of variability at the extremes of the distribution (the top 1-5%). The lack of mean differences in a number of studies may be overshadowing any secondary findings about variability. With this in mind, it could be suggested that another type of “file-drawer” problem is in effect, resulting in a lack of findings published in the literature to do with differences in variability, significant or otherwise.

Considering intelligence in a more general context, the current finding is further contradictory to the long-standing position held by many that males are more variable than females (Geary, 1996; Hedges & Nowell, 1995; Jensen, 1998). Yet, the conclusion that males have greater variance in mean scores is not universally accepted, with some claiming that there lacks strong evidence in the literature to make such claims (Feingold, 1992; Mackintosh, 1996; 2001). This position further corroborates the current findings.

Discussions of variability in the general population appear to be of tertiary importance if they are addressed at all. While there is very little written about the variability on the SPM, conclusions from other measures suggest that sex differences in mean and variance are generally small. It is, however, at the extremes of the distribution where larger differences can be detected (Nowell and Hedges, 1998). If there were found to be significantly more males at the extremes of the distribution, there would be greater proportions of males in the groups of the population's highest and lowest cognitively achieving individuals. Although not directly proven in the literature, anecdotally the argument could then be made that differences in variability in the extremes are related to the findings that there are proportionately more males in high-level mathematical and science programmes and careers. Further, there are proportionately more males than females referred for special educational support for such learning disabilities as dyslexia (Anderson, 1997; Rutter et al., 2004).

Due to sample size constraints and the resultant effect on power to detect significant differences, investigations of the extremes of the distribution were not possible in the current dissertation. However, they are noted as a potential opportunity for further understanding, and will be discussed in a subsequent section pertaining to future research directions (section 8.7).

8.4. Differences in Mean and Variability at Younger and Older Ages

Due to the large age range of participants of the standardisation sample, it was important to consider whether age was contributing to the emergence of sex differences in general intelligence as had been suggested in the literature (Lynn, 1994; 1999; 2002). This was the third aim of this dissertation: to assess whether sex differences emerged in younger and older groups of participants. Effectively, the third aim of this study can also be considered an assessment of the validity of Lynn's Developmental Theory of Sex Differences in a large representative population. This proves to make a novel contribution to a literature with particularly narrow scope and research participation.

As discussed previously, Lynn proposed a Developmental theory that suggests that sex differences in cognition emerge at different points along the developmental continuum for boys and girls. He cites this as a reason for the lack of significant findings in some studies. He contends that significant differences in intelligence in these studies are being masked by a failure to account for the maturational differences in the development of boys and girls. According to this theory, girls are thought to have an advantage of approximately 1 IQ point in the earlier years of development which lasts until about 15. At this point, their physical and cognitive maturation decelerates relative to males, who then are thought to outperform females on tests of general intelligence by an estimated 2.4 IQ points (Lynn, Allik, & Must, 2000).

In the current dissertation, there was a failure to find a significant difference in mean scores between males and females on the SPM+ sample as a whole. In order to address Lynn's suggestion that significant sex differences are masked by a failure to account for age in analyses of sex differences, development was accounted for in the current analyses by assessing whether differences existed in younger and older groups of participants separately. As detailed in section 7.2 measurement invariance was first assessed, and it was determined that the measurement properties were equivalent for males and females in both the younger and the older age groups. In other words, items on the SPM+ measure general intelligence in the same way for boys and girls in both the younger and older age groups. Upon this basis, it was considered appropriate to proceed with a comparison of means and variability in the younger and older age groups.

Using MG-CFA methods, four groups were analysed separately and simultaneously in order to allow for equivalence of measurement properties and item parameters to be assessed across the groups. Verification within-sex was first conducted. It was determined that older males and females performed significantly better than the younger males and females. Further, the variability of scores for the older participants was found to be equivalent to the younger participants. These analyses were served as verification of validity. The results for mean differences were as expected, however the results for variance were not as expected. Logically, one would anticipate that children who are older to be more cognitively developed having experienced more diverse intellectual opportunities, and would therefore obtain a greater range of higher scores on measures of intelligence.

Next, assessments were conducted within-age across-sex. Variance of the latent factors was constrained to equality for males and females in each age group in order to isolate means of each group to allow for direct comparison. When comparing the mean performance of boys and girls in the group of younger participants, girls achieved 0.01 standard deviations higher than boys, but not significantly so. When comparing the older males and females, the females achieved 0.026 standard deviations higher than boys. Again this was not a significant difference. Equality constraints were then imposed upon the model in order to isolate the variance so that group comparisons could be made. No significant differences in score variability were found between younger boys and girls, or between older boys and girls.

The current findings, based upon a representative U.K. sample, disconfirm previous findings of a developmental trend in the SPM. Lynn, Allik, and Must (2000) identified a female advantage in 12 and 13 year-old females and a male advantage at 17 years in a sample of 12 to 18 year-olds in Estonia. A further study of 12 to 18 year olds from Estonia illustrated a female advantage at 12 to 15 year olds. Males in this study were found to outperform females between 16 and 18 years (Lynn, Allik, Pullmann, & Laidra, 2004).

A meta-analysis conducted by Lynn and Irwing (2004) provided further evidence of a developmental theory in a review of 57 studies spanning participants six to 89 years of age. This meta-analysis provides analysis of studies of both the Standard and the Advanced Progressive Matrices (APM) as the authors felt that the APM measures the same non-verbal reasoning ability as the SPM. Across these measures, Lynn and

Irwing (2004) identified that boys obtain slightly higher means than girls from six to nine years of age, but not significantly so (*d* = 0.01 to 0.10). From 10 to 13 years, a higher non-significant mean emerges for females (*d* = -0.06 to 0.05). At 14 years of age, a male advantage emerges (*d* = 0.08) which, at 15 years, becomes significant and increases in effect size to (*d* = 0.10). By 18 years of age, the significant difference increases in size to 0.16d.

It is not explicitly clear from their review what the age range was for each of the measures they assessed. This would have been particularly useful information as, generally speaking, the SPM is suitable for individuals of general intellectual functioning (with current published norms available for seven to 18 year olds), while the APM are designed for individuals of higher levels of intellectual ability. It is therefore not clear whether the differences can be attributed to the latent trait of general intelligence or whether the effects of an advanced measure of intelligence were biased against some of the participants. Information about which participants took which tests would have added considerably to the general understanding of their specific research findings as well as to their generalisability in a larger context.

As mentioned in section 2.3.2.5, evidence of the developmental theory of sex differences on the SPM has only been provided by Lynn and colleagues. While there are a handful of studies showing a developmental element, they make use of measures that are thought to have lower *g* loadings than the Raven's matrices (DAT: Colom & Lynn, 2004; NNAT: Rojahn & Naglieri, 2006). In what is likely to be the first non-Lynn assessment of the Developmental Theory of Sex Differences using the SPM, this dissertation provides evidence (substantiated by Rohjahn & Naglieri, 2006) that age is not a significant element in the emergence of sex differences in intelligence.

It was not possible to expand upon the assessment of mean and variance in young and older groups to verify the performance of males and females at the extremes of the distribution on account of insufficient sample sizes in the two sexes and the two age groups. Further studies of the sex differences in variance at the top and bottom of the distribution on the SPM+ are identified as opportunities for further study and will be discussed further in section 8.7.

8.5. Influence of Method Effects

The final aim of this dissertation was to assess the possibility that some of the differential covariance of the indicators could be attributed to the existence of a method effect. Although not noted in the published literature on the SPM, it has been noted in other literature that a potential cause of erroneous conclusions about group differences is a method effect. A method effect exists when some of the differential covariance amongst a set of items is attributable to some other aspect of the measure than the underlying latent ability being measured (Brown, 2006). Method effects (also referred to as 'methods factors') can arise from the modality of measurement presentation (such as the way items are worded or presented) or even to the pressures of social desirability (Brown, 2006; Podsakoff et al., 2003). The resultant method effect is a measurement artefact of different response styles and is not based upon differences in underlying latent factor. At present, there are no studies of the Raven's Matrices that assess method effects that could be attributed to item solving strategies that have been found to differ between males and females.

For illustrative purposes, an example is available from the Self-Esteem Questionnaire (SEQ; Marsh, 1996), and was presented earlier in section 2.4. Marsh challenged the commonly-used two-factor solution of the SEQ through the use of a single factor solution with a method effect, and identified that the items were not based upon substantively different dimensions of self-esteem, but rather the covariation of the factor loadings was related to the positive and negative item wording. Within the context of the Raven's Matrices, the multi-dimensional factor structures previously identified could potentially be attributed to method effect rather than latent factors.

During the assessment of group differences in mean and variability for the younger and older age groups of the SPM+, it emerged from the one-factor solution that a number of the indicators did not have significant and/or salient factor loadings. Upon further inspection many of the items were found to be common across all four of the groups in the model. When these items were removed from the model, there was found to be a significant improvement in global model fit. This finding may suggest that some of the covariance among these items is related to something other than the latent factor, such as the method of item presentation or the item type. The possibility of a methods effect was assessed in sections 6.4 and 7.4.

In light of the current dissertation and the longstanding body of literature claiming a male advantage in general intelligence, it was considered vital to evaluate the role of method effects as a source of potential psychometric bias. To do so, methods factors were assessed by extending the two-group and the four-group models to include two methods factors. The methods factors were defined according to the results of the model in accordance with existing literature on the measurement properties of the SPM.

While many theorists would contend that the Raven's Matrices are among the best measure of the unidimensional construct of general intelligence, (Abad, Colom, Rebollo, & Escorial, 2004; Court, 1983; Jensen, 1998; Mackintosh, 1996), others would contend that the Matrices measure multiple aspects of cognitive functioning. Lynn, Allik, and Irwing, (2004) argued for a 3-factor structure: Gestalt Continuation, Verbal-analytic Reasoning, and Visuospatial Ability. Others posit a 2-factor structure (Carpenter, Just, & Shell, 1990; Deshon, Chan, & Weissbein, 1995; van der Ven & Ellis, 2000), where the Progressive Matrices measures 'g' and a second perceptual or spatial factor.

In the studies that look specifically at multiple factor structure of the SPM (Lynn, Allik, & Irwing, 2004; van der Ven & Ellis, 2000), what is common across these studies is that they advocate for a Gestalt factor in the earlier items of the test, while a number of later items are considered to be measuring Visuospatial elements of ability. A number of the residual variances in the two- and four-group models of the current analyses clustered together in patterns that were reminiscent of these multi-factor conceptions, suggesting methods factors that were related to a Gestalt and Visuospatial elements. Consequently, it was from the perspective of these multi-factor models that the residual variances from the models were viewed, and according to which the method factor models were ultimately specified.

For the two-group model of sex differences, residual variances from 19 indicators were loaded onto two methods factors: Gestalt and Visuospatial ability. When the effects of the factors were assessed, there was no difference in the effect of the Gestalt factor for males or females. However, males were found to be slightly negatively influenced by the Visuospatial nature of the items B6-12, C1, C3, and C6 (z = 2.224; *p* = 0.026). Next, the methods factors were applied to the four-group model that assessed sex and age simultaneously.

The four-group MG-CFA model with two methods factors (gestalt and visuospatial) would not converge for the younger group of males and females. It is likely that a model with such complexity was not supported by the data from the younger participants, and is likely related to the large number of freely estimated parameters in the model (Brown, 2006). The exact cause of inability of the model to converge is unclear from the information available from the non-converged model. It was therefore not possible to estimate a MG-CFA with two methods factors for younger males and females for the U.K. standardisation data of the SPM+.

For the older participants the measurement invariance model provided evidence that the influence of the methods factors is the same for older boys and girls. When each of the methods factors was isolated in succession, it was revealed that there was no significant difference in the mean of the first methods factor (or what might be thought of as Gestalt items) or in the mean of the second methods factor (or what might be thought of as Visuospatial items).

What the results from analysis of the residual variance by method factors suggests is that in addition to measuring *g*¸ there are elements of the SPM that are attributable to Gestalt and Visuospatial answering strategies. For the overall sample, the effect of the Gestalt factor was the same for males and females. However, the effects of the visuospatial factor were slightly detrimental only to the males of the sample.

When attempts were made to assess the effects of the answering strategies by methods factor in groups of younger and older participants, the model would not converge for the younger group. For the older group, no significant differences in the effects of the methods factors could be determined. In light of the significant effect of the visuospatial factor in the overall sample, the lack of effects in the older sample, and the lack of convergence in the younger sample, it could be suggested that the lack of convergence in the model of younger participants may relate to the kinds of item solving strategies employed by younger respondents.

As the literature of item solving strategies has previously been conducted with older individuals (the youngest was reported to be 12 years, Lynn, Allik & Must, 2000), it may be that existing conceptions of solving strategies and item types do not hold for younger participants in the same way. The answering strategies used by younger participants could have been involved in the significant effect of the Visuospatial factor for males in the overall sample. This is identified as a potential opportunity for further research investigation.

8.6. Summary of Findings

In addressing the four research objectives of this dissertation, a number of conclusions can be drawn. First, the SPM+ measures the construct of general intelligence in the same way for males and females in the sample as a whole, as well as in groups of younger and older participants. This provides evidence that the items of the SPM+ are not biased to either males or females, and provides an equitable assessment of their general cognitive functioning.

Secondly, when the whole sample of participants aged seven to 18 years of age are considered, males and females perform equally well in terms of mean score and variability of score distribution. This provides substantial evidence that males and females are more equal in terms of the latent construct of general intelligence than has been previously suggested in the literature (e.g., Irwing & Lynn, 2005; Lynn & Irwing, 2004; Rushton & Skuy, 2000).

Thirdly, when age was accounted for in the assessment of sex differences in the SPM+, again, males and females were found to perform equally well. This finding refutes Lynn’s Theory of Developmental Sex Differences that, until now, has only been empirically tested by Lynn and his colleagues.

Finally, analyses revealed that the SPM+ predominantly measures general intelligence, but that some of the residual variance can be explained by Gestalt- and Visuospatial-type answering strategies. When these method effects were tested for a sex difference, in the overall sample, there was evidence that males were significantly disadvantaged by the Visuospatial nature of some of the items. This difference disappeared for the older males once the age was accounted for in the analyses. However, it was not possible to assess this finding in relation to the younger participants due to the lack of model convergence.

In light of the findings of the four research aims, it is believed that prior findings of sex differences in general intelligence as measured by the SPM have been concluded in error. This error was likely the result of one or both of the following: the use of samples that were inadequately representative of the population at large; or the use of statistical procedures that did not allow for the assessment of item bias and the evaluation of group differences in the latent trait of general intelligence while accounting for measurement error. The current dissertation addressed both of these issues by employing Multiple-Groups Confirmatory Factor Analysis techniques with a sample representative of seven to 18 year old males and females in the U.K. A conclusion of no appreciable sex differences in general intelligence can be made confidently and soundly.

8.7. Strengths, Limitations, & Future Research Directions

There are two notable strengths of the current dissertation. The first was the quality, size and representative nature of the data sample. This sample was further strengthened by the method of assessment used: the SPM is a measure with a consistent and reliable empirical history. The existing literature of significant sex differences on the SPM makes strong claims that have largely been based upon opportunity samples. It could be argued that these findings do not generalise well to the population at large. It has been suggested that it is inappropriate, both ethically and psychometrically, to generalise to a wider population when convenience samples have been used (Hunt & Madhyastha, 2008). The size and representativeness of the current U.K. standardisation sample of the SPM+ is thought to be a novel contribution to the field of study of sex differences in intelligence. Based on the noted quality of the sample employed in this research, it is therefore considered appropriate to generalise the conclusions from this study to a wider population.

The second strength of this dissertation is that it made use of the most up-to-date statistical methodology for investigating group differences in latent constructs. To date, the large majority of the existing literature has employed classical statistical methods for assessing sex differences, such as t-tests and analysis of variance. With the advent of modern statistical modelling, such methods are no longer considered to have the strength and specificity required to assess the nuances at the item level that can translate into group differences in latent traits (Embretson & Reise, 2000).

Structural equation modelling methods, such Multiple Groups Confirmatory Factor Analysis used in this dissertation, allow for the equivalence of measurement properties to be assessed across groups. This ensures that any influences of bias at the item level, or differential item functioning, are accounted for in the model. Bias occurs when a test item yields a different response for members of different groups who have the same level of the latent trait. Further, modelling techniques ensure that differences, or lack thereof, are true and not artefacts arising from measurement error (Wicherts et al., 2005; van der Sluis, et al., 2008).

In the current dissertation, assessment of measurement invariance ensured that the SPM+ was measuring the construct of general intelligence in the same way for males and females, and as such, there was no bias inherent in the measure that would interfere with the performance of either group. It is upon the foundation of measurement invariance that conclusions regarding the lack of differences in SPM+ mean performance and variability can be drawn.

Despite these strengths, this dissertation encountered limitations worthy of mention. The most notable limitation of this dissertation was the identification of a number of items that proved to load non-significantly and non-saliently when the items were imposed upon one latent factor of general intelligence as discussed in section 5.5. Further inspection revealed that a number of the items in the one factor model showed low communalities with factor loadings that fell below a minimum threshold of 0.3. According to this guideline recommended by MacDonald (1999), items that do not meet this minimum criterion are not considered to be meaningfully representative of the factor upon which they are loaded.

Evidence from the Exploratory Factor Analysis conducted in Chapter 5 also revealed that a number of the items loaded strongly onto more than one factor, also known as cross-loading. It is advised (Brown, 2006; McDonald, 1999) that indicators that do not reach the minimum factor loading threshold of 0.3, or load upon multiple factors simultaneously should not be included in the analyses, unless there are theoretical grounds to do so.

In the current dissertation, it was felt that there were substantial grounds upon which to argue for the inclusion of the non-significant, non-salient, and cross loading items. Because the SPM+ is a published test of intelligence that is used extensively in empirical research, educational settings, and clinical practice, all test items were retained in the analyses of this dissertation for reasons of completeness, comparability to other published findings, and for generalisability to the population at large. An extension of the current research would be to re-analyse the data using a one-factor model after the removal of the unsuitable items.

One possible explanation of the unsuitable factor loadings may be related to the constructs of intelligence that are being measured by the Raven's Matrices. The Raven's Matrices were designed according to a one-factor model (Raven, 2009), and accordingly, a number of prominent theorists would maintain that the Raven's Matrices measure general intelligence, and virtually nothing else (Jensen, 1998; Raven, 2009; Raven, Raven, & Court, 1998c). However, it has often been suggested that, in addition to general intelligence, the Matrices measures additional constructs of intelligence. Lynn, Allik, and Irwing, (2004) argued for a 3-factor structure whereby the matrices measure three additional factors of gestalt continuation, verbal-analytic reasoning, and visuospatial ability. Others advocate a 2-factor structure (Lynn, Backhoff, & Contreras-Nino, 2004;

Carpenter, Just, & Shell, 1990; Deshon, Chan, & Weissbein, 1995; van der Ven & Ellis, 2000), where the SPM measures ‘g’ and a second perceptual or spatial factor.

In light of these multi-factor possibilities, the non-salient and non-significant loadings from the present one-factor model may have been due to the fact that the SPM+ is measuring more than general intelligence. A further direction in which to take the findings from the current dissertation would be to re-analyse the data using a multi-factor perspective. This would address a number of things. First, the multiple factor solution might account for the cross-loadings that were found in the current one-factor model more effectively. Further, it is possible that the strength of the factor loadings upon multiple factors would increase in order to satisfy the minimum criteria of significance and salience. Finally, approaching analyses of the SPM from the perspective of multiple factors using modelling techniques such as multiple-groups confirmatory factor analysis would make a contribution to the field that has yet to be made.

A further limitation encountered in the analyses of this dissertation was the inability to effectively assess the influence of methods factors in the younger age groups. When the methods factors were included in the model for the younger group of participants, the model failed to converge. As discussed, the failure of the model to converge was likely related to the large number of freely estimated parameters in the model (Brown, 2006). In light of the possibility that the SPM+ may be measuring multiple constructs of cognitive ability, one further direction in which to take the current research would be to verify the likelihood that the multiple factor structure would hold for younger and older participants. In so doing, it would be possible to address the question relating to differential item functioning and bias that is attributable to different answering strategies that may be employed by boys and girls at different ages. Further, the influence of methods effects in a multi-factorial model could be assessed.

The advantages of the strength and specificity of statistical modelling techniques (such as multiple-groups confirmatory factor analysis) is becoming ever apparent in the literature of general intelligence, and intelligence in general. The number of studies using these analytic methods is increasing in the literature, and it is hoped that such techniques will prove to balance the largely inconsistent findings reported thus far. Assessing the latent construct of intelligence as a function of the respondent's level of the latent construct in relation to the characteristics of the test items allows for more precise understanding of any differences in test scores than if assessed with classical methods. It is hoped that future endeavours to identify sex differences in intelligence will employ such measures to ensure accuracy and authenticity of their conclusions.

In accordance with the Gender Similarities hypothesis (Hyde, 2005) it is appropriate to say that males and females in the U.K. are more similar than they are different, with respect to their general intellectual ability, as evidenced by analyses of the standardisation sample of the SPM+. It is hoped that the results reported in this dissertation of no sex differences in performance on the SPM+ will provide some balance to what has been largely a one-sided debate arguing that men have superior general intelligence to women (Begley, 2009).

The debate concerning sex differences in intelligence has been argued for generations, and will likely continue to be argued for generations to come. An excerpt from an article in Nature (1923) poignantly illustrates that, in fact, the conclusions about similarities in general intellectual ability of the sexes have changed very little:

> "Sex is the cause of only a small fraction of the mental differences between individuals... It has been stated, upon statistical grounds, that the largest sex-differences are physical differences – differences in height, in weight, and in bodily strength. Intellectual differences are far smaller...in the higher and more complex processes – in general intelligence and in ability to reason – the differences during the school period are extremely small" (p. 658).

Despite this seemingly longstanding belief that males and females differ very little in terms of "mental differences", there remains a pervasive viewpoint in the literature of general intelligence that males are superior to females. However, the mere existence of a perspective does not necessarily mean it exists in reality. In order for a theory to be considered highly plausible, the theory must be supported with sound, empirical research conducted in a variety of settings, using different representative samples of participants, and using different measurement methods (Halpern, In Press). It could be argued that a highly plausible theory for male superiority in general intelligence has yet to be presented in the literature. Rather, the body of literature suggestive of a highly plausible theory of gender similarities in general cognitive ability is further strengthened by the findings of the present investigation.

APPENDIX 1:

PARENT INFORMATION LETTER

Project on children's problem solving and vocabulary abilities
University of Cambridge

Dear Parent/Guardian,

A project is being carried out across the United Kingdom involving the Raven's test, a well-known non-verbal test of problem solving ability and its complementary vocabulary test. The aim is to ensure that a large group of children is tested over the next few months. By testing a large number of children in many different schools it is possible to understand how children of each age typically perform. It will then be possible in future to assess the strengths and developmental needs of an individual child.

Your child's school is one of 120 throughout the country that are involved in this project, and we would be most grateful if you would give permission for your child to take part. Participation in this study is voluntary, but if you agree, your child may be tested at school either in a group or individually on a test of problem solving ability using patterns and a test of vocabulary by a qualified tester. The tests will take about 45 to 60 minutes and will involve your child in a series of paper and pencil tests. For example, your child will be asked to complete a series of shapes and patterns, and asked to give the meaning of some words. You or your child would be free to withdraw at any time and without giving a reason.

All information about your child will be kept strictly confidential. Your child's test results will not be given to the school and your child's name will not be attached to the results. If you are happy for your child to take part, please complete the enclosed Parental Consent Form. Children will then be randomly selected for testing. The information requested in the form is to ensure that we include children from a wide variety of backgrounds. If we have enough children from each group then we will not ask your child to take part in the project. You are, of course, under no obligation to agree to your child taking part in this project and your refusal will not affect your child's education or care in any way. If you do not wish your child to be tested please tick the "no" box on the Parental Consent Form and you will not be contacted again.

By taking part in this project your child will be helping to ensure that other children in the future receive the support they need at school. Your co-operation is very important to us and greatly appreciated.

Yours sincerely,

Professor John Rust

Information Sheet

If you consent to your child taking part in this project they may be tested at school with the Raven's non-verbal test and a vocabulary test.

The test results will be confidential. This means that:

- We will protect the confidentiality of the information you provide within the limitations of the law.
- Your personal details will be known only to the researcher in charge of the study and will be kept in a locked filing cabinet at Cambridge Assessment, Cambridge University.
- The test results will be held in a separate locked filing cabinet at Cambridge Assessment, Cambridge University with no identifying information attached. An identification number will be used in place of your child's name. Not even the researchers will be able to identify an individual child's test results.
- All of the information held in filing cabinets will be shredded in 18 months with the project ends.
- Information entered onto computer for analysis will be in the form of numbers.
- Your test results will be used for statistical purposes only.
- When the results of the project are published you will not be identified as having taken part in the research, neither will information which might make you identifiable be published.
- As the analyses will be carried out anonymously we shall not be able to provide an individual child's results.

If you have any queries please do not hesitate to contact Emily Savage-McGlynn at Cambridge Assessment on 01223 552 708.

APPENDIX 2:

PARENTAL CONSENT FORM

Name of parent/guardian: ______________________________________

Parent/guardian's address: ______________________________________

____

Child's last name: ____

Child's first name: ____

Name of child's school: ____

Child's sex: □ **Male** □ **Female**

Child's date of birth: _____/_____/_____ **(Day / Month/ Year)**

What is your child's race/ethnicity?

(*If they are of mixed race you my tick more than one)*

CRE classification

□ White
□ Black-Caribbean
□ Black-African
□ Black-Other
□ Chinese

□ Indian
□ Pakistani
□ Bangladeshi
□ Other (please specify)

Does the mother/female guardian live at home with the child?

□ **Yes** □ **No**

What is the mother's/female guardian's main job? __________________

Does the father/male guardian live at home with the child?

□ **Yes** □ **No**

What is the father's/ male guardian's main job? ____________________

Which of the following qualifications does the mother/female guardian

have?

- □ 1+ O levels/CSEs/GCSEs/ 1 + O-grade/standard grade (any grades)
- □ 5+ O levels/ 5+CSEs (grade 1) /5+ GCSEs (grades a-c), School Certificate/ 5 + O- grades/standard grades
- □ 1+ A levels/ AS levels/ Higher Grades/ Certificate of Year 6 Grade
- □ 2+ A levels, 4+ AS levels, Higher Grades/ Certificate of Year 6 Grade
- □ First Degree (e.g. B.A., B.Sc.)
- □ Higher Degree (e.g. M.A., Ph.D., PGCE, post-graduate certificates/ diplomas)
- □ NVQ Level 1, Foundation GNVQ
- □ NVQ Level 2, Intermediate GNVQ
- □ NVQ Level 3, Advanced GNVQ
- □ NVQ Levels 4-5, HNC, HND
- □ Other Qualifications (e.g. City and Guilds, RSA/OSR, BTEC/Edexcel)
- □ No Qualifications

Which of the following qualifications does the father/male guardian have?

- □ 1+ O levels/CSEs/GCSEs/ 1 + O-grade/standard grade (any grades)
- □ 5+ O levels/ 5+CSEs (grade 1) /5+ GCSEs (grades a-c), School Certificate/ 5 + O- grades/standard grades
- □ 1+ A levels/ AS levels/ Higher Grades/ Certificate of Year 6 Grade
- □ 2+ A levels, 4+ AS levels, Higher Grades/ Certificate of Year 6 Grade
- □ First Degree (e.g. B.A., B.Sc.)
- □ Higher Degree (e.g. M.A., Ph.D., PGCE, post-graduate certificates/ diplomas)
- □ NVQ Level 1, Foundation GNVQ
- □ NVQ Level 2, Intermediate GNVQ
- □ NVQ Level 3, Advanced GNVQ
- □ NVQ Levels 4-5, HNC, HND
- □ Other Qualifications (e.g. City and Guilds, RSA/OSR, BTEC/Edexcel)
- □ No Qualifications

Data Protection

I understand that any information I provide is confidential, and that no information that could lead to the identification of any individual will be disclosed in any reports on the project, or to any other party. No identifiable personal data will be published. The identifiable data will not be shared with any other organization.

I also understand that my participation is voluntary, that I can choose not to participate in part or all of the project, and that I can withdraw at any stage of the project without being penalized or disadvantaged in any way.

Do you agree for your child to be tested at school with the Raven's non-verbal test and its complementary vocabulary test?

□ **Yes** □ **No**

Signed: ____

Date: ____

PLEASE RETURN THIS FORM TO THE SCHOOL

APPENDIX 3:

RAVEN'S SPM+ ADMINISTRATION INSTRUCTIONS

To be read to the participant(s) verbatim:

"To begin with we are going to look at some patterns. Look at the Standard Progressive Matrices answer sheet. All of your answers should be made on this answer sheet.

Now take your test booklet. Please don't mark it in any way. Open your test booklet at the first page. You see that this is problem number A1. Now look at your answer sheet. You will see that under the heading Set A there is a column of numbers A1, A2, A3, A4, through to A12. This is where the answers go.

*Now look back at your test booklet. The top part of Problem A1 is a pattern with a bit cut out of it. Look at the pattern, and try to figure out which piece is needed to complete the pattern correctly both along and down. Then choose the right one out of the six pieces shown below. Each of these pieces below (***pointing to each in turn***) is the right shape to fill the space, but only one of them is the right pattern.*

*Number 1 is the right shape, but is not the right pattern. Number 2 is not a pattern at all. Number 3 is quite wrong. Number 6 is nearly right, but is wrong here. Only one is right. (***Give the students time to consider the answer options***).*

*Number 4 is the right bit isn't it? So the answer is Number 4. Find Set A on your answer sheet. Now put a single line through 4 next to A1 like this (***demonstrating on the example***).*

*Now please turn to the next page of your test booklet and do Problem A2 by yourselves. (***Give the students time to consider the answer options***).*

The right answer is Number 5. Have you put a line through '5' next to problem A2 on your answer sheet?

On every page of the booklet there is a pattern with a piece missing. You have to choose which of the pieces below is the right one to complete the pattern. When you think you have found the right piece, put a line through its number next to the problem number on your answer sheet. The problems are simple in the beginning and get harder as you go on.

If you make a mistake, or want to change your answer, put a cross through the incorrect answer and then put a single line through the number of the correct answer. Do not try to rub out the incorrect answer.

Go on like this by yourself until you get to the end of the booklet. Work at your own pace. There is no time limit. I will check to see that you are getting on alright.

Try not to miss any out. If you are not sure, guess, as guesses are sometimes right. If you get really stuck, move on to the next problem and then, if you want, come back to the one you had difficulty with. Any questions?

Turn to Problem A3 and start."

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.

Abdel-Khalek, A., & Lynn, R. (2006). Sex differences on the Standard Progressive Matrices and in educational attainment in Kuwait. *Personality and Individual Differences*, *40*, 175-182.

American Association Of University Women. (1993). *College Admissions Tests: Opportunities or Roadblocks?* Washington: Author.

Anderson, K. G. (1997). Gender bias and special education referrals. *Annals of Dyslexia*, *47*(1), 151-162.

Arden, R., & Plomin, R. (2006). Sex differences in variance of intelligence across childhood. *Personality and Individual Differences*, *41*(1), 39-48.

Baker, F. (2001). *Basics of Item Response Theory*. College Park: ERIC Clearinghouse on Assessment and Evaluation.

Barnett, R., & Rivers, C. (2005). *Same Difference*. New York: Basic Books.

Begley, S. (2009). Sex, Race and IQ: Off Limits? *Newsweek*, *153*(16), 53.

Bentler, P. M. (1990). Comparative Fit Indices in Structural Models. *Psychological Bulletin*, *107*, 238-246.

Binet, S. (1905). New Methods for Diagnosis of the Intellectual Level of Subnormals. *L'Année Psychologique*, *12*, 191-244.

Blinkhorn, S. (2005). A gender bender. *Nature*, *438*.

Block, J. H. (1976). Issues, problems, and pitfalls in assessing sex differences: A critical review of The psychology of sex differences. *Merrill-Palmer Quarterly*, *22*, 283-308.

Bors, D. A., & Stokes, T. L. (1998). Raven's Advanced Progressive Matrices: Norms for First-Year University Students and the Development of a

Short Form. *Educational and Psychological Measurement*, *58*(3), 382-398.

Bradley, R. H., & Corwyn, R. F. (2002). Socioeconomic Status and Child Development. *Annual Review of Psychology*, *53*, 371-399.

Brody, N. (1992). *Intelligence* (2nd ed.). London: Academic Press.

Brouwers, S. A, Van De Vijver, F. J. R., & Van Hemert, D. A. (2009). Variation in Raven's Progressive Matrices scores across time and place. *Learning and Individual Differences*, *19*, 330-338.

Brown, T. A. (2006). *Confirmatory Factor Analysis for Applied Research*. (D. A. Kenny, Ed.). London: The Guildford Press.

Browne, M. W., & Cudek, R. (1993). Single Sample Cross-Validation Indices for Covariance Structures. *Multivariate Behavioral Research*, *24*, 445-455.

Burt, S. C. (1955). *The Subnormal Mind*. London: Oxford University Press.

Byrne, B. M., Shavelson, R. J., & Muthén, B. O. (1989). Testing for the Equivalence of Factor Covariance and Mean Structures: The Issue of Partial Measurement Invariance. *Psychological Bulletin*, *105*, 456-466.

Cahill, L. (2006). Why sex matters for neuroscience. *Nature reviews. Neuroscience*, *7*(6), 477-84. Nature Publishing Group.

Camarata, S., & Woodcock, R. (2006). Sex differences in processing speed: Developmental effects in males and females. *Intelligence*, *34*(3), 231-252.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, California: Sage.

Caplan, P. J., MacPherson, G. M., & Tobin, P. (1985). Do Sex-Related Differences in Spatial Abilities Exist? *American Psychologist*, *40*, 786-799.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: a theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review*, *97*(3), 404-431.

Carroll, J. B. (1993). *Human Cognitive Abilities*. Cambridge: Cambridge University Press.

Cattell, R. B. (1970). *Abilities: Their structure, growth, and action.* Boston: Houghton Mifflin.

Cattell, R. B., & Horn, J. L. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology*, *57*(5), 253-70.

Ceci, S. J., Williams, W. M., & Barnett, S. M. (2009). Women's underrepresentation in science: sociocultural and biological considerations. *Psychological bulletin*, *135*(2), 218-61.

Cianciolo, A. T., & Sternberg, R. J. (2004). *Intelligence: A brief history.* Blackwell Publishing: Malden.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioural Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates Publishers.

Collins, N. (2010). Separate GCSEs for boys and girls. *The Daily Telegraph*, *June 18, 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.

Colom, R., & García-López, O. (2002). Sex differences in fluid intelligence among high school graduates. *Personality and Individual Differences*, *32*(3), 445-451.

Colom, R., Juan-Espinosa, M., Abad, F. J., & García, L. F. (2000). Negligible sex differences in general intelligence. *Intelligence*, *28*(1),57-68.

Colom, R., Lluis-Font, J. M., & Andres-Pueyo, A. (2005). The generational intelligence gains are caused by decreasing variance in the lower half of the distribution: Supporting evidence for the nutrition hypothesis. *Intelligence*, *33*, 83-91.

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.

Conlin, M. (2003). The New Gender Gap: From Kindergarten to Grad School, Boys are Becoming the Second Sex. *Business Week*, May 26.

Costa, D. I., Azambuja, L. S., Portuguez, M. W., & Costa, J. C. (2004). [Neuropsychological assessment in children]. *Jornal de pediatria*, *80*(2 Suppl), S111-6.

Costello, A. B., & Osborne, J. W. (2005). Best Practices in Exploratory Factor Analysis: Four Recommendations for Getting the Most From Your Analysis. *Practical Assessment, Research & Evaluation*, *10*(7), 1-9.

Court, J. H. (1983). Sex differences in performance on Raven's Progressive Matrices: A review. *Alberta Journal of Educational Research*, *29*, 54-74.

Crane, P. K., Belle, G. V., & Larson, E. B. (2004). Test bias in a cognitive test: differential item functioning in the CASI. *Statistics in Medicine*, *256*, 241-256.

Crane, P. K., Gibbons, L. E., Jolley, L., & Belle, G. van. (2006). Differential Item Functioning Analysis With Ordinal Logistic Regression Techniques. *Medical Care*, *44*(11), 115-123.

Crane, P. K., Narasimhalu, K., Hays, R. D., Cella, D., Gibbons, L. E., Ocepek-Welikson, K., et al. (2007). A comparison of three sets of criteria for determining the presence of differential item functioning using ordinal logistic regression. *Quality of Life Research*, *16*(0), 69-84.

Crocker, L., & Algina, J. (2006). *Introduction to Classical and Modern Test Theory*. New York: Wadsworth Publishing Co.

Crucian, G. P., & Berenbaum, S. A. (1998). Sex differences in right hemisphere tasks. *Brain and cognition*, *36*(3), 377-89.

Deary, I. J., Irwing, P., Der, G., & Bates, T. (2007). Brother–sister differences in the g factor in intelligence: Analysis of full, opposite-sex siblings from the NLSY1979. *Intelligence*, *35*(5), 451-456.

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and Educational Achievement. *Intelligence*, *35*, 13-21.

Deary, I. J., Thorpe, G., Wilson, V., Starr, J. M., & Whalley, L. J. (2003). Population sex differences in IQ at age 11: the Scottish mental survey 1932. *Intelligence*, *31*, 533-542.

Demetriou, A., Mouyi, A., & Spanoudis, G. (2008). Modelling the Structure and Development of g. *Intelligence*, *36*, 437-454.

Department for Education. (2010). *Statistical First Release National Curriculum Tests and Teacher Assessments at Key Stage 2 & 3 in England, 2010 (Provisional)*. London.

Der Sluis, S. van, Derom, C., Thiery, E., Bartels, M., Polderman, T. J. C. A., Verhulst, F. C., et al. (2008). Sex differences on the WISC-R in Belgium and the Netherlands. *Intelligence*, *36*, 48-67.

Deshon, R. P., Chan, D., & Weissbein, D. (1995). Verbal overshadowing effects on Raven's advanced progressive matrices: Evidence for multidimensional performance determinants. *Intelligence*, *21*(2), 135-155.

Deshon, R. P., Chan, D., & Weissbein, D. A. (1995). Verbal overshadowing effects on Raven's Advanced Progressive Matrices: Evidence for multidimensional performance determinants. *Intelligence*, 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 Measurement1*, *41*, 1295-1302.

Docherty, S. J., Kovas, Y., Petrill, S. A., & Plomin, R. (2010). Generalist genes analysis of DNA markers associated with mathematical ability and disability reveals shared influence across ages and abilities. *BMC Genetics*, *11*, 61-71.

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are the central issues. *Psychological Bulletin*, *95*(1), 134-135.

Dykiert, D., Gale, C., & Deary, I. J. (2009). Are apparent sex differences in mean IQ scores created in part by sample restriction and increased male variance? *Intelligence*, *37*(1), 42-47.

Désert, M., Préaux, M., & Jund, R. (2009). So young and already victims of stereotype threat: Socio-economic status and performance of 6 to 9 years old children on Raven's progressive matrices. *European Journal of Psychology of Education*, *24*(2), 207-218.

Eccles, J. S. (1994). Understanding women's educational and occupational choices: Applying the Eccles et al. model of achievement-related choices. *Psychology of Women Quarterly*, *18*, 585-609.

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, New Jersey: Lawrence Erlbaum Associates Publishers.

Eysenck, H. J. (1981). *Intelligence: the battle for the mind*. (H. J. Eysenck & L. J. Kamin, Eds.)*Intelligence: the battle for the mind*. London: Palgrave Macmillan.

Fan, X., Chen, M., & Matsumoto, A. (1997). Gender Differences in Mathematics Achievement: Findings From the National Education Longitudinal Study of 1988. *The Journal of Experimental Education*, *65*(3), 229-242.

Fanous, A., Gardner, C. O., Prescott, C. A., Cancro, R., & Kendler, K. S. (2002). Neuroticism, major depression and gender: a population-based twin study. *Psychological Medicine*, *32*, 719-728.

Feingold, A. (1992). Sex Differences in Variability in Intellectual Abilities: A New Look at an Old Controversy. *Review of Educational Research*, *62*(1), 61-84.

Feng, J., Spence, I., & Pratt, J. (2007). Playing an Action Video Game Reduces Gender Differences in Spatial Cognition. *Psychological Science*, *18*(10), 850-855.

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin*, *101*, 171-191.

Flynn, J. R. (2007). *What is Intelligence*. Cambridge: Cambridge University Press.

Flynn, J. R. (2009). Requiem for nutrition as the cause of IQ gains: Raven's gains in Britain 1938-2008. *Economics and Human Biology*, *7*(1), 18-27.

Galton, F. (1908). *Memories of my life*. London: Methuen.

Gardner, H. (1993). *Frames of Mind*. London: Fontana Press.

Geary, D. C. (1996). Response: A Biosocial Framework for Studying Cognitive Sex Differences. *Learning and Individual Differences*, *8*(1), 55-60.

Gottfredson, L. S. (1998). The General Intelligence Factor. *Scientific American Presents*, *9*, 24-29.

Gottfredson, L. S., & Saklofske, D. H. (2009). Intelligence: Foundations and issues in assessment. *Canadian Psychology/Psychologie canadienne*, *50*(3), 183-195.

Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin*, *103*(2), 265-75.

Gurian, M., & Stevens, K. (2005). *The Mind of Boys: Saving Our Sons from Falling Behind in School and Life*. San Francisco: Jossey-Bass.

Gustafsson, J.-E. (1984). A unifying model for the structure of intellectual abilities. *Intelligence*, *8*(3), 179-203.

Halpern, D. F. (2000). *Sex Differences in Cognitive Abilities* (3rd ed.). Mahwah, New Jersey: Lawrence Erlbaum Associates Publishers.

Halpern, D. F. (2004). A Cognitive-Process Taxonomy for Sex Differences in Cognitive Abilities. *Current Directions in Psychological Science*, *13*(4), 135-139.

Halpern, D. F. (2007). Science, Sex, and Good Sense: Why Women are Underrepresented in Some Areas of Science and Math. In S. J. Ceci & W. M. Williams (Eds.), *Why Aren't More Women in Science?* Washington: American Psychological Association.

Halpern, D. F. (In Press). *Sex Differences in Cognitive Abilities* (4th ed.). Mahwah, New Jersey: Lawrence Erlbaum Associates Publishers.

Halpern, D. F., & LaMay, M. (2000). The Smarter Sex: A Critical Review of Sex Differences in Intelligence. *Educational Psychology Review*, *12*(2), 229-246.

Hedges, L. V. (2008). What Are Effect Sizes and Why Do We Need Them? *Child Development Perspectives*, *2*(3), 167-171.

Hedges, L. V., & Nowell, A. (1995). Sex Differences in Mental Test Scores, Variability, and Numbers of High-Scoring Individuals. *Science*, *269*(5220), 41-45.

Hines, M. (1990). Gonadal Hormones and Human Cognitive Development. In J. Balthazart (Ed.), *Hormones, brain, and behavior in vertebrates. 1. Sexual differentiation, neuroanatomical aspects, neurotransmitters and neuropeptides.* (pp. 51-63). Basel: Karger.

Hines, M. (2005). *Brain Gender*. Oxford: Oxford University Press.

Hines, M. (2010). Sex-related variation in human behavior and the brain. *Trends in Cognitive Sciences*, *14*(10), 448-456.

Hohm, E., Jennen-Steinmetz, C., Schmidt, M. H., & Laucht, M. (2007).

Language development at ten months. Predictive of language outcome and school achievement ten years later? *European Journal of Child and Adolescent Psychiatry*, *16*(3), 149-56.

Homer. (2007). *The Odyssey*. Radford, VG: Wilder Publications.

Horgan, D. M. (1975). *Language Development.* Ann Arbor: University of Michigan.

Horn, J. L., & McArdle, J. J. (1992). A Practical and Theoretical Guide to Measurement Invariance in Aging Research. *Experimental Aging Research*, *18*(3), 117-144.

Howe, M. A. J. (2000). *IQ in Question: The Truth About Intelligence*. London: Sage.

Hu, L., & Bentler, P. M. (1999). Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives. *Structural Equation Modeling*, *6*, 1-55.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item Response Theory: Application to Psychological Measurement*. Homewood, IL: Dow Jones-Irwin.

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.

Hyde, J. S. (2005). The Gender Similarities Hypothesis. *American Psychologist*, *60*, 581-592.

Hyde, J. S. (2007). New Directions in the Study of Gender Similarities and Differences. *Current Directions in Psychological Science*, *16*, 259-263.

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender Differences in Mathematics Performance: A Meta-Analysis. *Psychological Bulletin*, *107*(2), 139-153.

Hyde, J. S., & Linn, M. C. (1988). Gender Differences in Verbal Ability: A

Meta-Analysis. *Psychological Bulletin*, *104*(1), 53-69.

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.

Irwing, P., & Lynn, R. (2006). Intelligence: is there a sex difference in IQ scores? *Nature*, *442*, E1-2.

Jackson, D., & Rushton, J. P. (2006). Males have greater g: Sex differences in general mental ability from 100,000 17- to 18-year-olds on the Scholastic Assessment Test. *Intelligence*, *34*(5), 479-486.

Jensen, A. R. (1998). *The G Factor: The science of mental ability*. Detroit, Michigan: Praeger.

Johnson, W. (2004). Just one g: consistent results from three test batteries. *Intelligence*, *32*(1), 95-107.

Johnson, W., Carothers, A., & Deary, I. J. (2008). Sex Differences in Variability in General Intelligence: A New Look at an Old Question. *Psychological Science*, *3*(6), 518-531.

Johnson, W., Carothers, A., & Deary, I. J. (2009). A role for the X chromosome in sex differences in variability in general intelligence. *Perspectives on Psychological Science*, *4*, 598-611.

Johnson, W., McGue, M., & Iacono, W. G. (2007). Socioeconomic Status and School Grades: Placing their Association in Broader Context in a Sample of Biological and Adoptive Families. *Intelligence*, *35*, 526-541.

Jones, R. (2009). *Item Response Theory: A Primer*. University of California, Davis.

Kashy, D. A., Donnellan, M. B., Ackerman, R. A., & Russell, D. W. (2009). Reporting and interpreting research in PSPB: practices, principles, and pragmatics. *Personality and Social Psychology Bulletin*, *35*(9), 1131-42.

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.

Khaleefa, O., & Lynn, R. (2008a). A Study of Intelligence in the United Arab Emirates. *Mankind Quarterly*, *49*(1), 58-64.

Khaleefa, O., & Lynn, R. (2008b). Sex Differences on the Progressive Matrices: Some Data from Syria. *Mankind Quarterly*, *48*(3), 345-351.

Kimura, D., & Hampson, E. (1994). Cognitive Pattern in Men and Women is Influenced by Fluctuations in Sex Hormones. *Current Directions in Psychological Science*, *3*(2), 57-61.

Lawton, C. A., & Morrin, K. A. (1999). Gender Differences in Pointing Accuracy in Computer-Simulated 3D Mazes. *Sex Roles*, *40*(1/2), 73-92.

Lietz, P. (2006). Issues in the change in gender differences in reading achievement in cross-national research studies since 1992: a meta-analytic view. *International Education Journal*, *7.2*, 127-149.

Lim, T. (1994). Gender-related differences in intelligence: Application of confirmatory factor analysis. *Intelligence*, *19*, 179-192.

Linn, M. C., & Petersen, A. C. (1985). Emergence and Characterization of Sex Differences in Spatial Ability: A Meta-Analysis. *Child Development*, *56*(6), 1479 - 1498.

Lippa, R. A, Collaer, M. L., & Peters, M. (2010). Sex differences in mental rotation and line angle judgments are positively associated with gender equality and economic development across 53 nations. *Archives of sexual behavior*, *39*(4), 990-7.

Logan, S., & Johnston, R. (2010). Investigating gender differences in reading. *Educational Review*, *62*(2), 175-187.

Lutchmaya, S., Baron-Cohen, S., & Raggatt, P. (2002). Foetal testosterone and eye contact in 12-month-old human infants. *Infant Behavior and*

Development, *25*(3), 327-335.

Lynn, R. (1990). The role of nutrition in secular increases of intelligence. *Personality and Individual Differences*, *11*, 273-285.

Lynn, R. (1994). Sex differences in intelligence and brain size: A paradox resolved. *Personality and Individual Differences*, *17*, 257-271.

Lynn, R. (1998). Sex differences in intelligence: data from a Scottish standardisation of the WAIS-R. *Personality and Individual Differences*, *24*, 289-290.

Lynn, R. (1999). Sex differences in intelligence and brain size: a developmental theory. *Intelligence*, *27*, 1-12.

Lynn, R. (2002). Sex differences on the progressive matrices among 15-16 year olds: Some data from South Africa. *Personality and Individual Differences*, *33*, 669-673.

Lynn, R., Allik, J., & Irwing, P. (2004). Sex differences on three factors identified in Raven's Standard Progressive Matrices. *Intelligence*, *32*, 411-424.

Lynn, R., Allik, J., & Must, O. (2000). Sex differences in brain size, stature and intelligence in children and adolescents: some evidence from Estonia. *Personality and Individual Differences*, *29*(3), 555-560.

Lynn, R., Allik, J., Pullmann, H., & Laidra, K. (2004). Sex differences on the progressive matrices among adolescents: some data from Estonia. *Personality and Individual Differences*, *36*, 1249-1255.

Lynn, R., Backhoff, E., & Contreras-Niño, L. (2004). Sex differences on g, reasoning and visualisation tested by the progressive matrices among 7–10 year olds: some normative data for Mexico. *Personality and Individual Differences*, *36*(4), 779-787.

Lynn, R., & Irwing, P. (2004). Sex differences on the progressive matrices: a meta-analysis. *Intelligence*, *32*, 481-498.

Lynn, R., & Mikk, J. (2009). National IQs predict educational attainment in math, reading and science across 56 nations. *Intelligence*, *37*, 305-310.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*(2), 130-149.

Maccoby, E. E., & Jacklin, C. N. (1974). *The psychology of sex differences.* Stanford, California: Stanford University Press.

Mackintosh, N. J. (1995). Insight into intelligence. *Nature*, *377*, 581-582.

Mackintosh, N. J. (1996). Sex Differences and IQ. *Journal of Biosocial Science*, *28* (Special Issue 04), 558-571.

Mackintosh, N. J. (2001). *I.Q. and Human Intelligence*. New York: Oxford University Press.

Mackintosh, N. J., & Bennett, E. S. (2005). What do Raven's Matrices measure? An analysis in terms of sex differences. *Intelligence*, *33*, 663-674.

Marsh, H. W. (1996). Positive and negative global self-esteem: a substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology*, *70*(4), 810-9.

Mau, W. C., & Lynn, R. (2000). Gender differences in homework and test scores in mathematics, reading and science at tenth and twelfth grade. *Psychology, Evolution, and Gender*, *2*, 119-125.

McDonald, R. P. (1999). *Test Theory: A Unified Treatment*. Mahwah, New Jersey: Lawrence Erlbaum Associates Publishers.

Mehl, M., Vazire, S., Ramírez-Esparza, N., Slatcher, R. B., & Pennebaker, J. W. (2007). Are Women Really More Talkative Than Men? *Science*, *317*(5834), 82.

Meredith, W. (1993). Measurement Invariance, Factor Analysis and Factorial Invariance. *Psychometrika*, *58*(4), 525-543.

Mohan, V., & Kumar, D. (1979). Performance of neurotics and stables on the standard progressive matrices. *Intelligence*, *3*, 355-367.

Molenaar, D., Dolan, C. V., & Wicherts, J. M. (2009). The power to detect sex differences in IQ test scores using Multi-Group Covariance and Means Structure Analyses. *Intelligence*, *37*, 396-404.

Moè, A. (2009). Are males always better than females in mental rotation? Exploring a gender belief explanation. *Learning and Individual Differences*, *19*(1), 21-27.

Muthén, L. K., & Muthén, B. O. (2009). Mplus statistical modeling software. Los Angeles,C.A. Muthén & Muthén.

National Statistics. (2009). *Annual Survey of Hours and Earnings*.

Nature. (1923). School and Sex: Report of the Consultative Committee on Differentiation of the Curriculum for Boys and Girls Respectively in Secondary Schools. *Nature*, *111*(2794), 657-658.

Neisser, U., Boodoo, G., Bouchard Jr., T. J., Boykin, A. W., Brody, N., Ceci, S. J., et al. (1996). Intelligence: Knowns and unknowns. *American Psychologist*, *51*(2), 77-101.

Niederle, M., & Vesterlund, L. (2007). Do Women Shy Away from Competition? Do Men Compete Too Much? *Quarterly Journal of Economics*, *122*(3), 1067-1101.

Nowell, A., & Hedges, L. V. (1998). Trends in Gender Differences in Academic Achievement from 1960 to 1994 : An Analysis of Differences in Mean, Variance, and Extreme Scores. *Sex Roles*, *39*(1/2), 21-43.

Onis, M. de, Onyango, A. W., Borghi, E., Siyam, A., Nishida, C., & Siekmann, J. (2007). Development of a WHO growth reference for school-aged children and adolescents. *Bulletin of the World Health Organization*, *85*(9), 660-667.

Oxford University Press. (1999). *Oxford English dictionary*.

Ozcaliskan, S., & Goldin-Meadow, S. (2005). Gesture is at the Cutting Edge of Early Language Development. *Cognition*, *96*, B101-B113.

Partchev, I. (2004). *A visual guide to item response theory* (e-book): Author.

Petrill, S. A. (1997). Molarity versus Modularity of Cognitive Functioning? A Behavioral Genetic Perspective. *Behavior Genetics*, *6*(4), 96-99.

Podsakoff, P. M., MacKenzie, S. B., Jeong-Yeon, L., & Podsakoff, N. P. (2003). Common Method Biases in Behavioral Research: A Critical Review of the Literature and Recommended Remedies. *Journal of Applied Psychology*, *88*(5), 879-903.

Rasch, G. (1966). An individualistic approach to item analysis. In P. F. Henry & L. N.W. (Eds.), *Readings in Mathematical Social Science*. Chicago: Science Research Associates.

Raven, J. (2000). The Raven's Progressive Matrices: Change and stability over culture and time. *Cognitive Psychology*, *41*, 1-48.

Raven, J. (2009). The Raven Progressive Matrices And Measuring Aptitude Constructs. *The International Journal of Educational and Psychological Assessment*, *2*, 2-38.

Raven, J. C., Court, J. H., & Raven, J. (2008). *Ravens Standard Progressive Matrices and Vocabulary Scales*. London: Pearson Assessment.

Raven, J., Raven, J. C., & Court, J. H. (1998a). *Raven's Advanced Progressive Matrices*. Oxford: Oxford Psychologists Press.

Raven, J., Raven, J. C., & Court, J. H. (1998b). *The Coloured Progressive Matrices and Vocabulary Scales.* Oxford: Oxford Psychologists Press.

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

Raven, J., Raven, J. C., & Court, J. H. (2008). *Raven's Coloured Progressive Matrices and Vocabulary Scales*. London: Pearson Assessment.

Rojahn, J., & Naglieri, J. A. (2006). Developmental gender differences on the Naglieri Nonverbal Ability Test in a nationally normed sample of 5–17 year olds. *Intelligence2*, *34*(3), 253-260.

Rushton, J. P. (1995). *Race, evolution and behaviour: A life history perspective.* New Brunswick, NJ: Transaction Publishing.

Rushton, J. P., & Cvorovic, J. (2009). Data on the Raven's Standard Progressive Matrices from four Serbian samples. *Personality and Individual Differences*, *46*(4), 483-486.

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

Rutter, M., Caspi, A., Fergusson, D., Horwood, L. J., Goodman, R., Maughan, B., et al. (2004). Sex differences in developmental reading disability: new findings from 4 epidemiological studies. *Journal of the American Medical Association*, *291*, 2007-2012.

Santor, D. A., Ramsay, J. O., & Zuroff, D. (1994). Nonparametric item analyses of the Beck Depression Inventory: Evaluating gender item bias and response option weights. *Psychological Assessment*, *6*, 255-270.

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.

Sharp, C., Goodyer, I. M., & Croudace, T. J. (2007). Biased mentalizing in children aged seven to eleven: latent class confirmation of response styles to social scenarios and associations with psychopathology. *Social Development*, *16*(1), 181-202.

Sharpe, D. (1997). Of Apples and Oranges, File Drawers, and Garbage: Why Validity Issues in Meta-Analysis Will Not Go Away. *Clinical Psychology Review*, *17*, 881-901.

Silverman, I., Choi, J., Mackewn, A., Fisher, M., Moro, J., & Olshansky, E. (2000). Evolved mechanisms underlying wayfinding: further studies on the hunter-gatherer theory of spatial sex differences. *Evolution and Human Behavior*, *21*, 201-213.

Spearman, C. (1927). *The abilities of man.* New York: Macmillan.

Stanley, J. C. (1993). Boys and girls who reason well mathematically. In G. R. Bock & K. Acrill (Eds.), *The Origin and Development of High Ability.* (pp. 119-138). New York: Wiley.

Steiger, J. H., & Lind, J. M. (1980). *Statistically Based Tests for the Number of Common Factors*. Paper presented at the annual meeting of the Psychometric Society, Iowa City.

Sternberg, R. J. (1982). *Handbook of Human Intelligence*. (R. J. Sternberg, Ed.). Cambridge: Cambridge University Press.

Sternberg, R. J. (1985). *Beyond IQ*. Cambridge: Cambridge University Press.

Sternberg, R. J. (1990). *Metaphors of Mind*. Cambridge: Cambridge University Press.

Sternberg, R. J., & Detterman, D. K. (1986). *What is intelligence? Contemporary viewpoints on its nature and definition.* Norwood, NJ: Ablex.

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.

Subrahmanyam, K., & Greenfield, P. (1994). Effect of Video Game Practice on Spatial Skills in Girls and Boys. *Journal of Applied Developmental Psychology*, *15*, 13-32.

Tach, L., & Farkas, G. (2006). Learning-Related Behaviors, Cognitive Skills, and Ability Grouping When Schooling Begins. *Social Science Research*, *35*(4), 1048-1079.

Terman, L. M. (1916). *The Measurement of Intelligence: An Explanation of and a Complete Guide for the use of the Stanford Revision and Extension of the Binet-Simon Intelligence Scale.* Boston: Houghton Mifflin.

Thurstone, L. L. (1931). *The reliability and validity of tests.* Ann Arbor: Edwards Brothers.

Thurstone, L. L., & Thurstone, T. G. (1941). *Factorial studies of intelligence.* Chicago: University of Chicago Press.

Tombaugh, T. N., Kozak, J., & Rees, L. (1999). Normative Data Stratified by Age and Education for Two Measures of Verbal Fluency: FAS and Animal Naming. *Archives of Clinical Neuropsychology*, *14*(2), 167-177.

Tucker, L. R., & Lewis, C. (1973). A Reliability Coefficient for Maximum Likelihood Factor Analysis. *Psychometrika*, *38*, 1-10.

Tzuriel, D., & Egozi, G. (2010). Gender differences in spatial ability of young children: the effects of training and processing strategies. *Child Development*, *81*(5), 1417-30.

van der Sluis, S., Posthuma, D., Dolan, C. V., Geus, E. J. C. de, Colom, R., & Boomsma, D. I. (2006). Sex differences on the Dutch WAIS-III. *Intelligence*, *34*, 273 - 289.

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.

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*(2), 250-70.

Webb, R. M., Lubinski, D., & Benbow, C. P. (2007). Spatial Ability: A Neglected Dimension in Talent Searches for Intellectually Precocious Youth. *Journal of Educational Psychology*, *99*, 397-420.

Wechsler, D. (1944). *The measurement of adult intelligence* (3rd ed.). Baltimore: Williams & Wilkins.

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale - Third Edition (WAIS-III)*. San Antonio: Pearson Assessment.

Wechsler, D. (2003). *Wechsler Intelligences Scale for Children - Fourth Edition (WISC-IV)*. San Antonio: Pearson Assessment.

Weiss, E., Kemmler, G., Deisenhammer, E., Fleischhacker, W., & Delazer, M. (2003). Sex differences in cognitive functions. *Personality and Individual Differences*, *35*(4), 863-875.

Wicherts, J. M. (2008). Book review of J. R. Flynn (2007): What is intelligence? Beyond the Flynn effect. *Netherlands Journal of Psychology*, *64*, 41-43.

Wicherts, J. M., Borsboom, D., & Dolan, C. V. (2010). Why national IQs do not support evolutionary theories of intelligence. *Personality and Individual Differences*, *48*, 91-96.

Wicherts, J. M., Dolan, C. V., & Hessen, D. J. (2005). Stereotype threat and group differences in test performance: a question of measurement invariance. *Journal of personality and social psychology*, *89*(5), 696-716.

Wurman, R. S. (2000). *Information Anxiety 2*. Toronto: QUE.

Wyse, A. E., & Mapuranga, R. (2009). Differential Item Functioning Analysis Using Rasch Item Information Functions. *International Journal of Testing*, *9*(4), 333-357.

Yu, C. Y. (2002). Evaluating Cutoff Criteria of Model Fit Indices for Latent Variable Models with Binary and Continuous Outcomes.