Development and Assessment of Computer-Game-Like Tests of Human Cognitive Abilities

Jason McPherson, BPsych (Hons), BA
School of Psychology, University of Adelaide

A thesis submitted in fulfillment of the requirements for the degree of
Doctor of Philosophy

June, 2008
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chapter 1</td>
<td>Overview and Literature Review</td>
<td>1</td>
</tr>
<tr>
<td>Chapter 2</td>
<td>Exegesis</td>
<td>33</td>
</tr>
<tr>
<td>Chapter 3</td>
<td>Published Paper 1: A speeded coding task using a computer-based mouse response</td>
<td>50</td>
</tr>
<tr>
<td>Chapter 4</td>
<td>Published Paper 2: Gs Invaders: Assessing a computer-game-like test of Processing Speed</td>
<td>71</td>
</tr>
<tr>
<td>Chapter 5</td>
<td>In Press Paper 3: Assessing the validity of computer-game-like tests of processing speed and working memory</td>
<td>99</td>
</tr>
<tr>
<td>Chapter 6</td>
<td>General Conclusion</td>
<td>144</td>
</tr>
<tr>
<td></td>
<td>References</td>
<td>159</td>
</tr>
<tr>
<td></td>
<td>Appendix A: Paper 1 Reprint</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Appendix B: Paper 2 Reprint</td>
<td></td>
</tr>
</tbody>
</table>
Summary

The present thesis describes the development and assessment of two computer-game-like tests designed to measure two cognitive abilities currently of considerable interest to many researchers: processing speed (Gs) and working memory (WM). It is hoped that such tests could provide a unique and important addition to the range of tests currently employed by researchers interested in these constructs. The results of five separate studies are presented across three published papers.

In Paper 1-Study 1 (N = 49) a speeded computerized coding test (Symbol Digit) using the mouse as the response device was assessed. Because speeded tests are thought to be highly sensitive to response methods (Mead & Drasgow, 1994) it was deemed important to first assess how a mouse response method might affect the underlying construct validity of a speeded coding test independently of whether it was game-like. Factor analytic results indicated that the computerized coding test loaded strongly on the same factor as paper-and-pencil measures of Gs.

For Paper 2-Study 1 (N = 68) a more computer-game-like version of Symbol Digit was developed, Space Code. Development of Space Code involved the provision of a cover story, the replacing of code symbols with ‘spaceship’ graphics, the situating of the test within an overall ‘spaceship cockpit’, and numerous other graphical and aural embellishments to the task. Factor analytic results indicated that Space Code loaded strongly on a Gs factor but also on a factor comprised of visuo-spatial (Gv) ability tests. This finding was further investigated in the subsequent study.

Paper 2-Study 2 (N = 74) involved a larger battery of ability marker tests and a range of additional computer-game-like elements were added to Space Code.
Space Code included a scoring system, a timer with additional voice synthesized countdowns, aversive feedback for errors, and background music. Factor analysis indicated that after a general factor was extracted Space Code loaded on the same factor as paper-and-pencil measures of Gs and did not load on a factor comprised of non-speeded Gv tests.

Paper 3-Study 1 (N = 74) was aimed at assessing a computer-game-like test of WM (Space Matrix) and further assessing Space Code within a broader network of tests. Space Matrix used a dual task format combining a simple version of Space Code with a visually presented memory task based on the Dot Matrix test (Miyake, Friedman, Rettinger, Shah, & Hegarty, 2001). The cover story and scoring system for Space Code was expanded to incorporate this additional memory element. Factor analysis indicated that Space Matrix was loaded on the same first order factor as standard WM tests and the Raven’s Advanced Progressive Matrices (Gf). Space Code was substantially loaded on the second order factor but was weakly loaded on each of two first order factors interpreted as Gs and WM/Gf.

A final study is presented (Paper 3-Study2) in which Space Code and Space Matrix was administered to a school aged sample (N=94). Space Matrix exhibited construct validity as well as predictive validity (as a predictor of school grades), while results for Space Code were less encouraging. Space Matrix and Raven’s Progressive Matrices showed comparable relationships to school grades for Mathematics, English and Science subjects.

It is concluded that the development of computer-game-like tests represents a promising new format for research and applied assessment of known cognitive abilities.
DECLARATION

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material which has been accepted for the award of any other degree or diploma of a university or other institute of higher learning, except where due acknowledgement is made in the body of the text.

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

The author acknowledges that copyright of published works contained within this thesis (as listed below) resides with the copyright holder(s) of those works.


Jason McPherson

Signed: __________________________    Date: ___27th June, 2008___
Acknowledgements

I am indebted regarding the broad idea for the thesis which was sparked by a suggestion made by Nick Burns in one of our many rambling but sincere research conversations. Nick, as my principal supervisor, co-author and friend, has supported my thinking, my lack of thinking, and nearly every stage of my academic development over the last eight years. I am indebted to him for supporting and guiding me in so many ways while also allowing and trusting me to develop the research program independently from the very beginning.

I am also thankful for the quality and support of my secondary supervisors. Michael Lee, who contributed to convincing me to do a PhD and who provided and trusted me sufficiently to invest in multimedia software and books that became integral to the technical development of software for the thesis. Ted Nettelbeck, who always had something encouraging, important and helpful to say to me.

I have also been greatly supported, much to their own detriment and frustration, by Mark Brown, Lynda Klopp, Geoff Matthews, Carmen Rayner and other staff members. No matter how many computers I broke, networks destroyed, or pieces of extra equipment I needed, I was always accommodated to get what I needed done with a smile and a mild ribbing.

There are a great number of others who helped and provided comments and feedback on the research program. Thanks to you all too.
Dedications

To my parents and my dearly departed brother.

To my friends in all their shapes and forms.
Chapter 1. Introduction and Literature Review

Overview

The research program underlying this thesis was aimed at developing and assessing two tests of human cognitive ability designed to look and feel like computer games. The research program was incremental in nature and lead to a series of five separate but related studies. The thesis is therefore centred upon these five studies which are reported across three published papers (presented as chapters in manuscript typeset here). These three papers are book-ended by chapters providing broader context and discussion relevant to the broader research program as a whole.

Chapter One, the present chapter, provides a summary of previous literature relevant to the use of computer games in psychological research and in particular to their use as assessment tools. Additionally literature related to the choice of ability constructs targeted for measurement in the present thesis is discussed before a more detailed outline of the empirical aims of the thesis are presented.

Chapter Two provides an exegesis (explanation) for each of the five studies. The aim of this chapter is to provide the reasoning for each study in the context of the broader research program. This chapter also provides a range of information that was outside of the scope of the published papers. Some of this information relates to methodological reasoning while some relates to extensive background development, programming, and piloting that was conducted. Additionally this chapter also includes some screenshots of tests used in the studies. Some of these were not presented in the original studies because of copyright restrictions and intellectual property agreements. A CD-ROM is also provided in the back cover of the thesis, with a folder for each paper, containing useable versions of the
software. Software files are named to match descriptions provided within the papers and a readme.txt file contains information related to using these.

Chapters Three, Four and Five contain the three published papers together with statements outlining my and my co-author’s respective contributions to the papers. The paper manuscripts have been reformatted to match the typeset for the thesis and the published proofs are presented in appendices. Chapter Six provides a summary of results and a concluding discussion based upon the outcome of the research program as a whole.

*Computers in Psychological Assessment*

Computers have been used for assessment purposes by psychologists for at least three decades, with uses ranging from the automation of simple true/false questionnaire items, to assistance in documenting complex psychological histories (Epstein & Klinkenberg, 2001). Computerisation of clinical instruments, personality scales, job attitude scales and various types of cognitive tests has lead to many advantages including: ease of administration and standardisation, less training for proctors, fewer scoring errors, faster (even immediate) results, fewer disposable materials and, in some cases, less opportunities for cheating (Epstein & Klinkenberg, 2001; Mead & Drasgow, 1994).

In the area of cognitive abilities testing the potential of these advantages has been a motivating force behind large scale projects such as the computerised batteries developed by the US military (Alderton, Wolfe, & Larson, 1997; Peterson et al., 1990; Ree & Carretta, 1998) and by various educational institutions (Bennett, 1998; Mead & Drasgow, 1994). Furthermore, many important advances in abilities testing, such as adaptive testing procedures, are dependent on tests
being computerised (Embretson, 2003). Computerisation is thus likely to play a major role in abilities testing in the 21st century.

Computers in Cognitive Abilities Assessment

Generally, most computerised ability testing has involved the computerisation of traditional paper-and-pencil (PP) tests. Research suggests that adequately designed computerised tests can be as valid as the traditional tests they are based on (Bugbee Jr., 1996; Epstein & Klinkenberg, 2001; Mazzeo & Harvey, 1988; Mead & Drasgow, 1994; Neumann & Baydoun, 1998; Van de Vijver & Harsveld, 1994). In their meta-analysis Mead and Drasgow (1994) estimated the cross modal correlation between traditional and computerised test versions to be about .72 for speeded tests and .97 for power tests. Speeded tests are defined as tests comprising items of minimal difficulty with individual performance differences primarily due to the speed with which items are completed. Power tests are defined as tests comprising items of substantive difficulty with individual performance differences primarily defined by accuracy of responses. Thus, although computerised tests, and especially speeded tests, may not be strictly equivalent to their traditional counterparts, some studies have provided further optimism regarding their psychometric properties.

Neuman and Baydoun (1998,) for example, administered a battery of 10 paper-and-pencil clerical ability tests and a computerised battery of the same tests to 411 undergraduate students, from two large North American universities, who were also employed as clerical workers within their institution. The batteries were administered four weeks apart with order varied so that half the sample received each test mode first. The cross modal correlations for the tests ranged from .79 to
.92 with the average being .87. Structural Equation Modeling indicated that the two test modes had equivalent construct validity; a single factor model fit the data very well (AGFI = .98, RMSR = .012). Additionally, they also collected clerical supervisor ratings for the students and found no difference in the predictive validity of the two test modes.

Many of the advantages of implementing traditional tests on computers appear limited to the added conveniences for the researcher, clinician, or institution administering the test. However, it is equally, or potentially more, important to consider how computerisation could benefit those who are taking the tests. Given that modern computers are capable of dynamic and interactive presentations, it is unduly limiting to simply implement via computer traditional tests that were never designed with computerisation in mind. Indeed, various military institutions have utilized computer technologies to develop dynamic assessment tools such as psychomotor tracking tasks and time estimation tasks with some apparent success (e.g. see McHenry, Hough, Toquam, Hanson, & Ashworth, 1990; Peterson et al., 1990; Ree & Carretta, 1998).

Peterson et al. (1990) reported on the development of a predictor battery to supplement the widely known Armed Forces Vocational Aptitude Battery (ASVAB). With a focus on testing, what they termed, perceptual-psychomotor skills they developed a highly customized response pedestal after finding keyboard and joystick response devices were difficult to use and calibrate. The development phase took approximately three years. McHenry, Hough, Toquam, Hanson and Ashworth (1990) subsequently reported on the predictive ability of various composite scores derived from this. Their analyses indicated that the perceptual-psychomotor composites, together with spatial ability, predicted soldier proficiency measures almost as well as the entire ASVAB.
However, these tests appear to have had little life outside of their military
assessment roles and were developed in a very different era of computer
technology, often using somewhat specialized equipment. If we are to bring
computerised testing into the 21st century we may do well to look towards what is
now among the most prevalent formats for interacting with computers, the
computer game

Prevalence of Computer Games

There are numerous definitions for computer game and this situation is
complicated by the existence of a number of not quite synonymous terms such as
video game, arcade game, and console game. These terms reflect the variety of
media that can, and have, been used to instantiate what might be more broadly
defined as electronic multimedia games. The term ‘computer game’ is probably
now more commonly used due to the rise in availability of personal home
computers. Indeed, many games originally designed for arcade or console
machines are now available as computer games. For the purposes of the present
thesis, the term computer game will be used to denote any electronic multimedia
product that is presented or described as being a ‘game’ by its publisher or
creator.

Surveys conducted to assess the prevalence of computer game play have
investigated the home computer game play of 816 six-to-eleven year olds in the
UK. They found that 77% of the sample played computer games and 24% played
games every day. Looking at game play durations, 75% reported playing for more
than half an hour and 14% reported playing for over two hours in a typical session.
Griffiths and Hunt (1995) similarly surveyed 387 adolescents, aged between 12
and 16 years, and found that 31% played computer games daily, 15% played four to five times a week, and 19% played more than once a week. Looking at play durations, they found that 10% played for more than four hours in a sitting, 26% played for two-to-four hours, and 23% played for at least half an hour. Yelland and Lloyd (2001) interviewed 934 Australian children aged between 10 and 13 years. Similarly to Griffiths and Hunt (1995), they found that 30% of the sample reported playing computer games on a daily basis. Looking at estimated game play in more detail, they found that 7% played 21-40 hours per week, 21% played 7-14 hours per week and 67% played up to 7 hours each week. In the USA, a national survey of 527 parents of 2-to-17 year olds (Gentile & Walsh, 2002) indicated the average time children spent playing computer games each day was one hour. The parents of 13-17 year old children reported an average of about 13 hours per week of game play. Clearly, children are gaining substantial exposure to computers through computer game play. Reflecting on the prevalence of computer game play amongst children, Yelland and Lloyd suggested that educators who fail to utilise computer technology effectively “are in danger of losing credibility as professionals in the information age” (2001, p. 177). It may be beneficial for psychologists interested in computerised assessment to be similarly aware of this possibility.

**Computer Games in Psychological Research**

In his 2003 Presidential Address to the American Psychonomic Society, David Washburn provided a brief history of computer games in psychological research (Washburn, 2003). He charted the rise of computer game research from a single paper published in 1972 until a plateau of about 62 papers per year was reached in 1994, which level continued to the time of his review. Washburn pointed towards computer games as providing a unique format for psychological research and
highlighted a number of important issues, as well as some of the advantages and disadvantages in using them. He began by recalling the occasion of his first computer purchase when the first thing he did was begin playing a game, his professor then arriving to exclaim “You’re not being . . . academic” (p. 185). This incident served as a catalyst to Washburn’s attempts over the next two decades to make the study of computer games scientific.

Perhaps in order to avoid a lengthy debate as to what might and might not be considered a computer game, Washburn stressed that the boundaries between computer based tasks and computer games is somewhat blurred and that this situation is further clouded by the “increasing complexity and verisimilitude of computer based simulations, virtual environments, and so forth” (p. 185). However, he concluded that “fundamentally, you know a game when you are playing it, and one person’s game (implying fun) might well be another person’s task (implying work) or chore (implying tedium)” (p. 185). We will proceed with an understanding that there is probably a continuum between tasks and games. Regardless of our definitions, there are a number of more pragmatic problems associated with computer game research that Washburn has pointed out.

*Computer Games in Psychological Research: Problems*

The first problem a researcher will likely come across when attempting to incorporate computer games in their research is the programming demands required to implement them. One may avoid this by using commercially available games but these will not provide the researcher with any control over which variables, if any, are recorded or the nature of the tasks involved. The alternatives are either to modify existing games, which is often made impossible by commercial operators who wish to protect intellectual property in the software; or to create one’s own games. But, as Washburn (Washburn, 2003, p. 186) points
out, developing computer games requires a “much more substantial investment” than creating paper-and-pencil tests or even more traditional computerised tasks (e.g., reaction time tasks, serial-probe tasks, etc.).

The second issue raised by Washburn is that computer games may introduce uncontrolled complexities to the experimental situation such as increased psychomotor and visual demands as well as motivational effects. Obviously, the effect of such complexities could also be positive but if these elements alter a task to the extent that the variables or underlying constructs of interest are no longer reliably measured, then the researcher will not have gained anything by all of their efforts.

The final problem Washburn raised is the possibility that research using computer games could be perceived as unserious or trivial because games are generally associated with entertainment and fun. This will probably not be a problem for researchers investigating the psychological effects of playing computer games, such as the determinants of game playing arousal (Schneider, 2004) or the effects of violent games on teenagers (Gentile, Lynch, Linder, & Walsh, 2004), but may be more of an issue for researchers aiming to use computer games as experimental tools to investigate processes underlying behaviour. It is for this reason that he suggests using a different term such as game-like task to describe “any serious cognitive and comparative research tests that employ the graphics, hardware, competition, or other elements of computer games” (p. 187). Hereafter in the present thesis, the term computer-game-like will be used to describe tasks or tests designed by researchers and incorporating such elements and computer, video or console game will be used where a commercially available game is referenced.
In light of these three potential problems described by Washburn, researchers interested in using computer-game-like tasks need to be prepared for a potentially expensive and time consuming development phase followed by a careful research phase to address any potential extraneous effects introduced with the computer-game-like format and the potential perception of the work as unserious. These issues pose challenges to the researcher but they are not necessarily an insurmountable barrier and there are a number of advantages that may be gained for the trouble. Washburn drew attention to four potential advantages in the use of computer-game-like tasks in psychological research.

Computer Games in Psychological Research: Benefits

The first advantage Washburn highlighted was the common platform computer-game-like tasks may provide. A good example of this comes from the Learning Strategies Program (Donchin, 1995) in which researchers from different laboratories used the same computer-game-like test, known as Space Fortress (Mane & Donchin, 1989), to study different aspects of skill acquisition. The goal of Space Fortress is to destroy enemy ships while avoiding being hit oneself and to eventually destroy the enemy’s main fortress. There are some 50 parameters or so that the experimenter can manipulate, such as the speed of hostile ships and the number of stimuli to be remembered. Participants receive continuous performance feedback via a visible scoring system. Space Fortress has been successfully used by researchers interested in various processes associated with skill acquisition, including attention, learning, memory, psychomotor skills, and intelligence (Donchin, 1995; Mane & Donchin, 1989; Rabbitt, Banerji, & Szymanski, 1989; Washburn, 2003).

The second advantage Washburn pointed to is the potential for computer-game-like tasks to elicit greater motivation from and thus better performance by
participants. To illustrate this he described a simple experiment where the same continuous performance task was given to two groups. The task involved a stream of letters flashed on the screen, of which 80% were the letter X and 20% were the letter H. Participants in one group were told to click the mouse button whenever H appeared but to ignore the X’s. They were instructed to be as quick and accurate as possible. Participants in the other group were told that they were playing a game in which the goal was to detect enemy ships (H’s), and ignore friendly ships (X’s), on radar. They were similarly told to perform as quickly and accurately as possible. However, despite the software being identical, participants given the game-like cover story were 12% faster in responding to the target H’s but made more commission errors, that is, they responded to the non-target X’s more frequently. The participants given the game-like cover story appeared to respond in a manner characterised by competition effects due to their effort to ‘win’ at the game (Washburn, Hopkins, & Rumbaugh, 1989). Washburn did not address the issue that whether performance can be considered to be improved in the game condition depends critically on whether one uses speed or accuracy of responding as the dependent measure. However, although the effects of computer-game-like tasks on motivation, attention and strategy are not yet very well understood, further evidence that computer-game-like elements can aid performance comes from research demonstrating that very simple game-like qualities, such as asking people to chase and catch learning stimuli, can improve learning and retention of lexigrams (Romski & Sevcik, 1996), classroom-relevant material (Washburn, Putney & Henderson, 1995, as cited in Washburn, 2003) and job skills (Washburn & Raby, 2003, as cited in Washburn, 2003).

Enjoyment and wellbeing are the next benefits of computer-game-like tasks posited in Washburn’s address. Washburn believes that this will be the case
because participants are likely to experience computer-game-like tasks as fun and enjoyable. He stated:

Games may make it possible to conduct research with children, with the aged, and with institutionalized populations, providing access to key psychological data while simultaneously improving their psychological fitness and happiness. (p. 190)

However, the only empirical evidence provided to support this possibility is drawn from non-human primate research. Washburn and Rumbaugh (1992) found that rhesus monkeys regularly chose computer-game-like tasks over a range of other activities (including toys) playing for over nine hours each day on average. In one experimental manipulation, monkeys chose to play computer-game-like tasks to receive food pellets rather than freely receiving food pellets for half an hour but without access to the games. They also found that computer-game-like tasks were more effective than toys (puzzles, boxes, swings) or social pairing (placing compatible sets of monkeys in the same cage) in reducing stress and maladaptive behaviors such as stereotypy, over-grooming, and aggression in the monkeys. Thus, computer-game-like tasks appear to be beneficial and desirable for non-human primates, in certain circumstances.

The final advantage addressed by Washburn is the possibility for computer game technology to deliver new opportunities for psychological research while also providing new tools to help answer old questions. For example, the use of virtual environments to mimic real world performance, such as complex social interactions (Blascovich, Loomis, & Beall, 2002) or air traffic control (Ackerman & Ciancolo, 2000; Ackerman & Kanfer, 1993), could lead to detailed study of behaviours that were previously difficult to access within the laboratory. However,
such investigations are likely to bear upon traditional concepts such as perception, attention, learning and problem solving.

One area that could gain much from the potential advantages of computer game technology is the study of individual differences in cognitive abilities.

*Computer Games and Cognitive Abilities*

Researchers investigating individual differences in cognitive abilities are primarily interested in assessing the peak performance of individuals. Logically then, a format that combines all of the conveniences of computerised testing with the potential to increase the motivation and enjoyment of testees would surely be appealing. Testing situations are known to be a source of anxiety for many people, especially children (McDonald, 2001). If computer-game-like tests are experienced as enjoyable and less like tests than are traditional assessment tools, they may evoke less anxiety in some individuals and thus provide a useful alternative, or supplemental assessment option. Furthermore, research investigating the structure of cognitive abilities often necessitates large test batteries, meaning participants often spend many hours at computer terminals. If computer-game-like tests prove to be more enjoyable and motivating than traditional computerised tests, then the fatigue and boredom effects could be reduced or eliminated. A test format that reduced anxiety and decreased boredom and fatigue could potentially provide a better indication of an individual’s peak performance, or true ability.

However, of the 70 papers published in 2001, the last year of Washburn’s review (2003), only two were assessment related. Indeed, there has been little attention paid to the potential for using computer games or computer-game-like tests in the assessment of cognitive abilities. The extant literature is now reviewed.
In one of the first studies I was able to identify examining computer game performance from a psychometric perspective, Jones, Kennedy and Bittner Jr. (1981) assessed the inter-trial stability of a commercially available console game, Atari’s Air Combat Maneuver. They tested 22 participants (Navy enlisted men) over 15 days, with 10 games played each day and they used the average number of enemy ‘hits’ as the dependent measure. They found that the standard deviation of scores stabilised after six days as did the inter-trial correlations. However, they did not assess the relationship between game performance and any external measures. Jones and colleagues (Jones, Dunlap, & Bilodeau, 1986) subsequently addressed this issue using a range of commercially available console games and traditional cognitive ability tests. This study therefore represents the first identifiable attempt to locate computer game performance within the measurement space of traditional cognitive ability tests.

Jones et al. (1986) administered five commercially available console games to 63 male psychology students, including the Air Combat Maneuver game discussed above, and 13 paper-and-pencil ability tests. Twelve of the paper-and-pencil tests were taken from the Educational Testing Service (ETS) Kit of Factor Referenced Cognitive Tests (Ekstrom, French, & Harman, 1976). The 12 ETS tests were chosen to measure verbal, quantitative, spatial, and perceptual abilities. The final test was a reversed printing task which was included as a measure of psychomotor ability. Participants played each of the five games 12 times with their average game score for each game used as the dependent measure for that game. These game scores were significantly correlated with all of the ability tests except the verbal tests. Although the magnitude of correlations was quite moderate (maximum $r = .32$, average $r = .19$) this may be partially due to the low reliability of game scores; the test-retest reliability of these was not high (range =
.57 to .63). Average multiple correlations for all of the tests ranged from .18 to .50. These results suggest that commercially available games share variance with traditional abilities tests but they are certainly not good measures of the same abilities constructs. Subsequent research comparing Air Combat Maneuver performance with a range of computerised cognitive and information processing tests found correlations of similar magnitudes (Kennedy, Jones, & Baltzley, 1988, as cited in Bliss, Kennedy, Turnage, & Dunlap, 1991). However, research comparing computer game and tracking task performance has indicated much stronger relationships (Bliss et al., 1991).

Bliss et al. (1991) used three of the same commercially available console games as Jones et al. (1986) as well as three conventional tracking tasks, commonly used in military assessment, with a sample of 40 male navy enlistees. Using scores obtained after task performance had stabilized, the average intercorrelation amongst the three tracking tasks \((r = .51)\) was about the same as amongst the three games \((r = .53)\). By comparison, and despite the games being administered one year after the tracking tasks, the average correlation between the computer games and the tracking tasks was .66. The authors suggested that video games could provide an attractive alternative to the traditional tasks for a number of reasons. They state that console games are commercially available and thus cheaper than the traditional tests which require specialised equipment.

Another notable attempt to assess the relationship between computer-game-like task performance and cognitive abilities utilized the Space Fortress platform described above (Mane & Donchin, 1989). Rabbitt, Banerji and Szymanski (1989) were interested in the relationship between IQ scores and psychomotor abilities such as those thought to be tapped by reaction time tasks. They were also interested in whether traditional paper-and-pencil measures would predict learning
rates of more complex interactive tasks such as Space Fortress. For this reason they were interested in assessing the relationship between traditional IQ scores and Space Fortress across multiple testing sessions. They also pointed to the advantages of computer-game-like tests such as Space Fortress, which can provide performance measures for subtasks as well as for overall performance, because this allows the relationships between task components and other tests to be more fully understood in relation to overall performance.

Rabbitt et al. (1989) tested 56 males in five one-hour sessions, consisting of five Space Fortress trials, with each participant thus playing Space Fortress a total of 25 times. Participants were also administered the Alice Heim 4 (Heim & Batts, 1948) test of general intelligence and 43 high scoring participants additionally completed the more advanced Alice Heim 5 (Heim, 1947) test of general intelligence. Participants were also asked to report the extent of their previous computer game experience. Results indicated that initial performance of Space Fortress was correlated moderately with previous game experience ($r = .33$), with AH 4 scores ($r = .28$), and with AH 5 scores ($r = .25$). Correlations with learning slope functions for Space Fortress performance across 25 administrations indicated that AH 4 ($r = .42$) and AH 5 ($r = .33$) scores were significant predictors of learning functions but previous computer game experience was not. Most participants obtained their maximum Space Fortress score on the final day of testing and maximum scores were strongly correlated with AH 4 scores ($r = .68$) but only moderately with AH 5 scores ($r = .30$). The smaller relationship with the AH 5 is somewhat perplexing given the test was designed to better differentiate high performers. It is possible that the AH 5 had poorer psychometric properties but this is not discussed by the authors. Previous computer game experience was
not a significant predictor of maximal game performance. In discussing their findings the authors’ state:

One interesting implication is that a relatively unsophisticated video-game, on which performance may reasonably be expected to be independent of native language or acquired literacy, and which is greatly enjoyed by young people who play it, rank orders individual differences in ‘intelligence’ nearly as well as pencil and paper psychometric tests which have been specially developed for this purpose over the last 80 years” (p.255)

Although one should keep in mind the relatively small sample size, the fact that the relationship with the AH 5 sample was somewhat smaller, and that the relationship was with substantially practiced performance of Space Fortress, this study nonetheless provides some grounds for optimism concerning the psychometric viability of computer-game-like assessment tools.

Further optimism regarding computer-game-like tests of cognitive abilities comes from a study conducted by Jackson III, Vernon and Jackson (1993). Jackson III et al. administered a computer-game-like tracking test as well as the paper-and-pencil Multidimensional Aptitude Battery (MAB). The computer-game-like test, Dynamic Spatial Task, involved shooting at moving spaceship-like targets at the top of the screen, using a joystick to move and fire from a base at the bottom of the screen. The difficulty of the task was increased by increasing the horizontal speed of the targets, introducing vertical movement to the horizontal movement of the targets, adding barriers between base and targets, adding movement to the barriers, and adding distraction by having the targets fire at the base. Presumably, these elements also added to the game-like feel of the task. The Dynamic Spatial Task was found to be internally reliable; intercorrelations
between four trial blocks ranged from .75 to .81 and each trial block was correlated .91 with total game scores, the latter being inflated part-whole correlations.

The MAB is structured similarly to the Wechsler scales but is designed for group testing, consisting of quite similar subtests, and full scale scores from the MAB and Wechsler are correlated very highly, $r = .91$ (Jackson, 1984, as cited in Jackson III et al., 1993). Factor analysis of the MAB scores in the study by Jackson III et al. resulted in two factors corresponding to the traditional Verbal and Performance distinction used in the Wechsler scales. Results indicated that scores on the Dynamic Spatial Task loaded on the Performance factor ($r = .51$), with this loading being larger than some of the Performance subtest loadings on this factor, but did not load on the Verbal factor ($r = - .07$)). These results showed that the computer-game-like test was a valid indicator of what has traditionally been known as Performance IQ. These results also echo those obtained by Jones et al. (1986), who similarly found a small negative relationship between commercially available computer games and tests of verbal ability.

However, there is an important point of contrast between the computer-game-like test used by Jackson III et al. (1993) and those used in the other computer game studies discussed previously. In the study conducted by Jackson III et al. scores for the computer-game-like test were obtained from a single administration of the test. In contrast, Jones et al. (1986) used average scores obtained from 12 administrations of five commercially available games. Bliss et al. (1991) averaged scores obtained only after task stability was achieved for both the commercially available games and the tracking tasks they used. Rabbitt et al. (1989) found a moderate relationship between IQ and initial performance of the Space Fortress Game but the strongest relationships with IQ was with the maximum scores from 25 administrations. This contrast reflects an important difference between two
somewhat different approaches to the study of computer-game-like tasks as assessment tools.

The commercially available tasks used by Jones et al. (1986) and Bliss et al. (1991) were complex simulation style games designed to prove challenging over many hours of gameplay. Similarly, the Space Fortress game platform used by Rabbitt et al. (1989) was designed as a complex simulation, presented in a computer-game-like format, for the study of complex skill acquisition (Mane & Donchin, 1989). In a somewhat related research program, some researchers have found relationships between traditional reasoning tests and computer simulations of complex dynamic problem solving situations (Gonzalez, Thomas, & Vanyukov, 2005; Kröner, Plass, & Leutner, 2005).

Kröner, Plass and Leutner (2005) administered Multiflux (Kröner, 2001) and the reasoning scale tests from the Berlin Intelligence Scale (BIS-K Jäger, Süß, & Beauducel, 1997) to 101 German High School students. Multiflux simulates a fictitious machine that must be adjusted via dials to optimize its performance. Participants are not given instructions but must learn how to achieve this via trial and error. The authors reported a correlation between Multiflux and BIS-K scores of .65. However, the administration of the task took 30 minutes and required a lengthy exploration phase (about 10 minutes) before the assessment phase. Additionally, complex simulations have often been reported as having poor reliability (Funke, 1995; Kröner et al., 2005) and low relationships with traditional intelligence measures (Leutner, 2002; Wittmann & Süß, 1999).

In contrast to complex simulation tasks, the Dynamic Spatial Task used by Jackson III et al. (1993) was specifically designed as a simple tracking task embedded in a computer-game-like format. Like traditional ability tests, this task required almost no training and was administered only once. However, despite its
simplicity the test successfully incorporated a number of game-like elements without compromising reliability and it shared substantial variance with a range of traditional paper-and-pencil measures of cognitive abilities. A similar approach has been taken by other researchers but the computer-game-like tasks developed have not been validated against traditional ability measures.

For example, Graham et al. (1985) reported the development of a computer-game-like measure of cognitive functioning called the Strategic and Tactical Assessment Record (STAR). The rationale behind this tool was to embed traditional psychological performance variables (e.g., reaction time, memory) within a larger game structure. The authors also pointed out that STAR was “specifically designed to have the strong motivational properties necessary in studies of sustained operations” (p. 643). STAR utilizes a ‘space war’ cover story in which the player is sent on missions that differ in difficulty and duration and in which approximately 80 performance measures are embedded. These measures fall into a smaller number of categories defined by the authors: psychomotor skill, perceptual accuracy and speed, and memory function. However, although the authors report experiments examining the effects of certain task parameters on performance, STAR performance measures were not compared with traditional tests. No further research with STAR appears to have been published.

Using a much simpler format than STAR, Ryan (1994) reported the development of a computer-game-like test of working memory, Memory for Goblins, based on the commonly used counting span paradigm. Memory for Goblins requires players to count the number of goblins presented on successive screens while holding the counts in memory to be reported later. The increasing set sizes are depicted as levels within the game. However, as with the STAR project (Graham et al., 1985), game performance was not compared with any
traditional measures. Furthermore, the original counting span task was minimally altered with no additional game-like elements (e.g., scoring, sound, feedback) beyond the use of ‘goblin’ graphics. However, even with this simple manipulation, the authors claimed that participants found the game “novel and interesting” (p. 215).

The possibility of making experimental tasks more interesting and motivating also appears to have prompted Berger, Jones, Rothbart, and Posner (2000) to develop a battery of computer-game-like tests of attention specifically for children. The authors suggest that although methodologies exist for measuring attention in toddlers and preschool children, they are only feasible if researchers “succeed in engaging the child’s attention and as long as the child stays on task” (p. 298). Because tests of attention developed for adults often involve hundreds of trials using rather boring stimuli (e.g., Test Of Variables of Attention; Leark, Dupuy, Greenberg, Corman, & Kindschi, 1999), the authors argue that they are unsuitable for very young children. To address this problem the authors developed four computer-game-like tests based on paradigmatic tests from the adult literature. These tests incorporated storylines in which the child was situated as an active agent, interacting with animals and other familiar objects via a touch-screen. In a sample of 30 children (aged between 58 and 62 months) the authors reported a range of attentional effects matching those often observed in adult studies using traditional attention tests. The authors also reported that children made very few errors, suggesting they were engaged in the tasks. However, because conventional abilities tests were deemed unsuitable for young children, it was not possible to compare game performance with any other psychometric criteria.
Aims of the Present Thesis

Researchers have used computer games in educational (Jayakanthan, 2002), experimental (Case, 1995; F. A. Gonzalez & Cathcart, 1995), clinical (Houghton et al., 2004; Lawrence et al., 2002), and surgical ability studies (Enochsson et al., 2004; Rosser et al., 2004). Notable amongst these studies, in a sample of 33 surgeons Rosser et al. found that computer game performance was a better predictor of simulated laproscopic surgery than the number of laproscopic cases performed or years of medical training. Research has also been conducted assessing the effects of computer game play on spatial ability (Sims & Mayer, 2002), information processing (Yuji, 1996), mental rotation ability (De Lisi & Wolford, 2002), gender differences in mental rotation ability (Quaiser-Pohl, Geiser, & Lehmann, 2006), visual search efficiency (Castel, Pratt, & Drummond, 2005), and even the development of theoretical analysis and reflective thinking skills (Aliya, 2002). Findings in this area have generally indicated positive affects for computer-game-play but there seem to be some limits to the generalisability of these affects (Sims & Mayer, 2002).

Despite these promising signs, there appears to be no published research specifically assessing the construct validity of computer games or computer-game-like assessment tools since the study conducted by Jackson III et al. (1993). The present thesis therefore aimed to address a gap in the extant literature by developing computer-game-like tests and assessing their construct validity. The development of these tests was guided by: issues arising from previous research; pragmatic considerations; and current theories and trends in cognitive ability assessment. These three areas of consideration are discussed below and the specific implications for the aims of the thesis are set out.
Issues Arising from Previous Research

One important issue evident in previous research (Bliss et al., 1991; Jones et al., 1986; Rabbitt et al., 1989) is that complex computer games require multiple administrations to provide a valid measure of traditionally measured abilities. Although this may be theoretically interesting, and relevant to the study of skill acquisition, it is clearly not ideal for an assessment tool in the context of research or applied purposes. Thus, the development of computer-game-like ability tests may prove more successful if games are comprised of simple subtasks. Such an approach was taken by Jackson III et al. (1993) and resulted in a computer-game-like task requiring only a single administration performance on which shared substantial variance with a range of traditional cognitive ability measures. However, this study highlights some other important issues in the development of computer-game-like ability tests.

Specifically, Jackson III et al. (1993) developed a computer-game-like test based on dynamic tracking tasks because previous research had suggested that such tasks were quite different from the static visual tasks used in paper-and-pencil test batteries. The authors were partially motivated by the possibility that computer-game-like tests might measure new constructs. In contrast, other researchers have aimed to develop computer-game-like tests to measure the same constructs as traditional ability tests. For example, the STAR task (Graham et al., 1985) utilized subtasks designed to provide well established cognitive performance indicators such as reaction times and memory spans. Similarly, Berger et al. (2000) sought to develop computer-game-like tests for children to measure well established attentional constructs identified within the adult literature. Although these studies did not assess the construct validity of the computer-game-like tasks, it is plausible that creating game-like tests involving
similar mental operations to traditional tests is likely to result in strong relationships between the two.

There would appear to be no *a priori* reason to prefer the development of computer-game-like tests to measure new abilities as opposed to established abilities. The discovery of new ability constructs may serve to expand our knowledge of intellectual functioning and provide further predictive power beyond traditional measures. However, as a first step, there are a number of pragmatic advantages to developing computer-game-like tests of established ability constructs. Developing more stimulating and enjoyable tests of established abilities could further our understanding of factors such as motivation, attention, and anxiety in psychological assessment. For example, at present it is difficult to assess the relative impact of test anxiety on ability measurement without tests that are less contaminated by anxiety to provide a baseline (McDonald, 2001). Furthermore, because a substantial literature already exists around the predictive validity of established ability constructs, computer-game-like tests of these abilities may prove more immediately interpretable in applied settings once their construct validity has been established. Finally, by attempting to develop computer-game-like tests of established ability constructs, we stand to gain knowledge even if we fail. If we go out of our way to specifically design a computer-game-like task to measure the same ability measured by a traditional test and fail, this will provide evidence that there are different or additional underlying abilities involved in computer-game-like tasks.

A final important issue raised in previous research is the role that previous game experience may have. Rabbitt et al. (1989) found a correlation between initial performance on Space Fortress and prior gaming experience but this was not a significant predictor of practiced performance or of learning slope functions.
for the game across multiple administrations. Most research using computer-game-like tasks has not addressed this issue but Jackson et al. (1993), on acknowledging this potential confound, suggested that this would be an important issue to address in future research. This may be of particular importance for computer-game-like tasks designed to be administered only once, because there will be no opportunity for those less familiar with games to practice the unique psychomotor skills that game-like tasks may require. It is also worth considering that it may be impossible to design computer-game-like tasks that will not be affected by prior computer game experience. However, the important issue is whether computer-game-like tests can be designed such that the greater proportion of test variance will be associated with the ability of interest and not with prior game experience.

In view of the above considerations, the present thesis focused on the development and assessment of relatively simple computer-game-like tests designed to measure well established constructs from the domain of known cognitive abilities. This was attempted by first using established tests as the basis for the computer-game-like tests that were developed, and subsequently using established tests as validity criteria for assessing the computer-game-like tests. It was thought that this would ensure that the computer-game-like tests were simple enough to provide meaningful performance measures from a single administration. Additionally, it was hoped that through this approach, and with careful consideration of the psychomotor demands of the computer-game-like instantiations, that the role of previous game experience would not interfere in the measurement of the established ability constructs of interest. The role of prior computer game experience was assessed, along with other variables known to be, or potentially, associated with computer game play and computer skills. For
example, gender has been shown to impact patterns and preferences for computer game play with males playing computer games more frequently (Phillips et al., 1995; Yelland & Lloyd, 2001) and showing a stronger preference for certain types of games (e.g., shooting games; Wood, Griffiths, Chappell, & Davies, 2004). Age effects were assessed because developmental trends in abilities are well documented (Horn & Blankson, 2005; Horn & Noll, 1997) and such effects should also be prevalent in computer-game-like tests. Despite previous research showing no differences between right and left handed people in right handed mouse skills (Hoffman, Chang, & Yim, 1997), handedness was also assessed for completeness.

**Pragmatic Issues**

Perhaps the most salient pragmatic difficulty in conducting research using computer-game-like tasks is the actual development of the software required. As Washburn pointed out, such tasks require substantially more time and resources than programming simple computerised tasks. For this reason, much of the research using such tasks has been developed within military contexts, including the Space Fortress program which was largely funded by military organizations (Shebilske et al., 2005). Furthermore, many of the computer-game-like tasks developed so far have utilized extremely simple graphics and sounds when judged by modern standards. Space Fortress for instance, although being structurally complex, utilizes rather simple stick like graphics. A recent upgrade to the Space Fortress platform opted to leave this aspect unchanged, despite it being obviously outdated (Shebilske et al., 2005). The computer-game-like tasks developed by Jackson III et al. (1993), Graham et al. (1985), Berger et al. (2000), and Ryan (1994) all utilized similarly simple graphics and sounds. However, a recent survey by Wood, Griffiths, Chappell and Davies (2004), investigating what they called the
'psychostructural' characteristics of computer games, found that realistic graphics and sounds were viewed as the most important game elements by self identified game players (2004). Importantly, this survey also identified a number of the other characteristics perceived as important by game players; for example a rapid absorption rate, character development, customizable game aspects, and multiplayer features. However, what does seem clear is that with the graphical and sound capabilities of computers rapidly increasing, researchers aiming to develop computer-game-like tests will presumably be confronted with increasing expectations for what constitutes a viable computer game.

Thus, the obvious pragmatic problem is that development of computer-game-like tests will demand a substantial investment of time and resources just to develop the game-like tasks ready for research. In the present thesis this development was necessarily managed by the author without external professional involvement. The development of the computer-game-like tests proposed for the present studies were aided by a focus on ability constructs that make item generation and manipulation reasonably manageable. It was also thought that the use of modern multimedia software would make the manipulation of graphics and sound feasible in the development process. However, the development of computer-game-like tests will require successive stages to deal with measurement issues arising from response modality differences, and within the restraints of a doctoral dissertation will be limited by obvious time and resource constraints. Nevertheless, the present thesis aimed to develop computer-game-like tests incorporating realistic graphics and sounds and as many other game-like elements, such as those identified by Wood et al. (2004), as possible.
Current Theories and Trends in Cognitive Ability Assessment

The idea of designing a computer-game-like test of cognitive ability leads to the inevitable question of which ability, or abilities, to measure. Although some previous research has compared computer game performance with global IQ scores (Rabbitt et al., 1989), researchers using multiple indicators of cognitive ability have found stronger relationships with tests of a more visual nature as opposed to verbal tests (Jackson III et al., 1993; Jones et al., 1986). However, the categorization of tests as verbal or visual is now regarded by many researchers as too simplistic. Although the entire history of intelligence testing has been characterized by a tension between those emphasizing the importance of a single underlying intelligence construct (e.g., Jensen, 1998) and those emphasizing various numbers of related or independent underlying ability constructs (e.g., Horn & Noll, 1997), there appears now to be some important points of convergence in regards to the general structure of known cognitive abilities (Carroll, 2005; Horn & Blankson, 2005; McGrew, 2005; McGrew & Flanagan, 1998). Arguably, the clearest explication of this convergence can be found in what has become known as the Cattell-Horn-Carroll (CHC) model of cognitive abilities (McGrew, 2005). Extensive detail regarding the evidence for the main structural elements presented in the CHC model have been provided in numerous reviews and summaries (Bickley, Keith, & Wolfle, 1995; Carroll, 1993, 2005; Flanagan & Harrison, 2005; Horn & Blankson, 2005; McGrew, 2005; McGrew & Flanagan, 1998; Roberts, Stankov, Pallier, & Dolph, 1997; Woodcock, McGrew, & Mather, 2001).

The CHC model represents an amalgam of what came to be known as Gf-Gc theory (Horn & Blankson, 2005; Horn & Noll, 1997) and Carroll’s Three Stratum theory (Carroll, 1993, 2005). Each of these theories provides a model detailing the structure of known human cognitive abilities. The CHC model appears to have
been developed primarily to present the elements common to these earlier theories, but using a single nomenclature to allow easier communication amongst and between researchers and clinicians (McGrew, 2005). Perhaps the primary difference between the two earlier models is that Gf-Gc theory does not include a higher level intelligence construct \((g)\) at the apex of all abilities, whereas Carroll’s theory does. The CHC model incorporates an overarching \(g\) construct, as in the Carroll model, as well as the two lower levels of organization common to both earlier models; the three organizational levels are referred to as strata. The first stratum is made up of over 60 abilities similar in scope to the abilities originally described by Thurstone (1938; 1947) as Primary Mental Abilities. Factor analyses of correlations amongst these first stratum abilities have lead to the postulation of about ten broad abilities at the second stratum. To reflect the fact that these second stratum abilities are also often found to be inter-correlated, the CHC model includes the third-stratum \(g\) factor. Fortunately, the debate surrounding both the existence and interpretation of this third stratum factor is relatively unimportant for the purposes of the present thesis. The reason for this is now explained.

Within CHC theory, individual tests are primarily classified according to their observed relationships with first and second stratum ability constructs (McGrew & Flanagan, 1998). Thus, in using the CHC model as a guiding taxonomy for the development of computer-game-like tests, the most pertinent question is which first or second stratum abilities might be most suitable for computer-game-like assessment. Answering this question, in turn, leads to a number of further questions of a rather pragmatic nature. Which abilities are most amenable to measurement using the types of stimuli commonly used in computer games? Which abilities involve the types of mental operations people are used to experiencing in computer games? Which abilities utilize question formats that are
most readily adaptable to game-like structures? There has been a growing trend within ability assessment research that points to two potentially suitable ability constructs that are also of great interest to researchers and clinicians alike.

Since the 1970’s there have been numerous attempts to explain variation in general intelligence measures by reference to more ‘basic cognitive processes’ (see Deary, 2001; Schweizer, 2005, for recent reviews). Two of the most frequently investigated processes, processing speed (Gs) and working memory (WM), are identified within the CHC model. Gs is measured by tests requiring rapid scanning and processing of visual information (Carroll, 1993) and is represented within CHC as a second stratum factor (Danthiir, Roberts, Schulze, & Wilhelm, 2005; Stankov & Roberts, 1997). WM is predominantly measured by tests requiring the shifting of attention back and forth between a memory span task and another, often quite simple, task (Engle, Tuholski, Laughlin, & Conway, 1999). However, WM has a perhaps contentious position within CHC taxonomies where it is generally depicted as a first stratum ability subsumed by a broad short term memory factor (Gsm; McGrew, 2005; McGrew, 1998). This position is contentious because a voluminous amount of research has indicated substantial relationships between WM and broader intelligence measures (Ackerman, Beier, & Boyle, 2005; Colom, Flores-Mendoza, & Rebollo, 2003; Conway, Kane, & Engle, 2003; Engle et al., 1999; Konig, Buhner, & Murling, 2005; Kyllonen & Christal, 1990; Mackintosh & Bennett, 2003; Schweizer & Moosbrugger, 2004). This issue is beyond the scope of the present thesis but the development of computer-game-like tests may provide another useful tool in this expanding research program. Of more direct import to the present thesis is the fact that a number of researchers have investigated the joint role of Gs and WM in intelligence (Conway, Cowan, Bunting, Therriault, & Minkoff, 2002; Fry & Hale, 1996, 2000; McGrew, 2005). Gs and WM
measures are also increasingly being used in more applied research (Colom, 
Escorial, Shih, & Privado, 2007; Swanson & Kim, 2007).

Recent research has identified Gs and WM as important interactive processes 
associated with cognitive deficiencies observed in people with multiple sclerosis 
(Lengenfelder et al., 2006). Similarly, Gs and WM measures have been found to 
predict cognitive deficits in children with myelomeningocele and shunted 
hydrocephalus even after controlling for age, math skill, and general intellectual 
functioning (Boyer, Yeates, & Enrile, 2006). Using data from the Woodcock- 
Johnson III Cognitive Abilities and Achievement Tests normative sample and the 
Western Reserve Twin Project, Luo, Thompson, and Dettermann (2006) found 
that combined Gs and WM scores correlated with scholastic achievement just as 
strongly as a combined measure of crystallized and fluid intelligence; the 
constructs generally thought to have the highest loadings on a general factor of 
intelligence (Carroll, 1993). However, as the authors of this study point out, these 
types of ‘process’ measures tend to be rigidly presented and appear ‘impersonal’ 
and therefore may be confounded by psychological effects such as anxiety and 
weakened motivation to a greater extent than other tests. There would appear to 
be considerable advantage in developing computer-game-like tests of Gs and WM 
constructs. There are a number of features inherent in the structure of these types 
of measures that also make them particularly suitable for game-like tests. These 
are now outlined.

Gs and WM are different from most other ability measures because they are 
thought to require minimal formal instruction and only very basic cognitive 
processes such as sequencing and classification (Luo et al., 2006). Both Gs and 
WM measures can be adapted to incorporate verbal, numerical, or spatial stimuli 
(Luo et al., 2006). This means that computer-game-like tests could be developed
without having to compromise the type of stimuli appropriate for this format. Additionally, these types of measures, being derived from the cognitive-experimental tradition, are extremely amenable to experimental manipulation (Luo et al., 2006; Schweizer, 2005). Such readily manipulated tasks ‘may facilitate the design and application of test items’ (Luo et al., 2006, p. 81). Considering the nature of computer games, it could also be postulated that these two constructs are already inherent features of commercial computer games. Many of these games involve speeded responding to stimuli, such as shooting enemies that appear unpredictably, a situation not dissimilar to many experimental speed tasks. Computer games also involve performance of simultaneous tasks to accomplish goals, including memory of objects seen or monitoring vital statistics of a character while performing other tasks, situations that could plausibly involve working memory.

The present research program thus proposed to develop computer-game-like tests of Gs and WM. It also proposed to make use of the extensive literature surrounding the CHC model of cognitive abilities by using established tests from the literature where appropriate. This guided selection of criterion tests was used to assess the construct validity of the newly developed computer-game-like tests. This approach is consistent with the cross-battery assessment methodology proposed by McGrew and Flanagan (1998). The nomenclature of CHC theory was adopted in the present thesis, to aid communication and interpretation of results, but the scope of experimental data did not extend to addressing the validity of the complete model. However, specific issues surrounding each of the proposed ability constructs are addressed in the relevant papers. Additionally, and in line with Luo et al.’s (2006) finding that information processing tasks were good predictors of scholastic performance, this thesis also aimed to assess the
predictive validity of the computer-game-like tests within two school aged samples using school grades as criteria.

An exegesis is now presented prior to three papers, the first two already being published and the third now accepted for publication. The exegesis is aimed at providing additional background and contextualizing information related to the studies described in the papers and explicating the relationships of each study to the broader research program underlying the present thesis. The exegesis also provides information relating to the development of the computer-game-like tests, and other computer based tests, that was outside the scope of the journal papers. Additionally, the exegesis also provides screenshots from some of the tests developed that could not be included in journal papers for reasons of copyright and intellectual property agreements.
Preamble

The broad rationale for this thesis originated from a recurring observation in various research projects with school and university student samples. Children, adolescents, and young adults participating in our studies would often express excitement, even relief, when told they would be using a computer. However, as screen after screen of static test items were presented, this initial enthusiasm was almost invariably seen to disappear. “This isn’t a game”, was the most common exclamation. It seemed to me that many of these younger participants associated computers with computer games and, although subsequent efforts were made to incorporate colourful graphics and sounds into our test programs, it became clear that we were not really fooling any of our participants. I had often observed participants, especially school aged children, lose motivation throughout a testing session whether using traditional or computerised methods. The implications of this concerned me, particularly in group testing situations where an individual administrator could not monitor any given individual’s level of task engagement. First, we were simply not providing our participants with an enjoyable experience with psychological research; and, second, it was unclear to what extent motivational effects were impacting our results. That is, if some children were differentially affected by boredom throughout the sessions this could greatly alter our results. Furthermore, if this boredom affected a certain type of child, or group of children, this bias might even be systematic. A question thus arose: Could computer games, which many children, adolescents, and adults, regularly use voluntarily for hours with apparent enjoyment, provide comparable data to that which traditional ability tests provided?
Examining the literature there were some positive signs, as described in the Introduction to this thesis, but it also became clear that I would have to develop my own customised games. First, there were no modern commercially available games that had been designed to or identified as measuring any specific ability construct. Second, previous research suggested that where commercially available games had been found to have some relationship with traditional measures, this was only after considerable practice on the computer game. Third, in order to be able to manipulate or fine tune an assessment tool it seemed necessary to have a degree of control over the tasks beyond that which commercially available games allow. It was clear I would need to be self sufficient in developing any computer-game-like assessment tools for my research program.

Thus, a crucial decision was on what software platform to use for the development of the computer-game-like tests. Based on research by Wood et al. (2004), I was aware that game players regarded realistic graphics and sounds, as well as scoring and feedback systems, as important elements in computer games. Interestingly, these were also the types of elements generally not used in computerised versions of traditional cognitive abilities tests. I needed to find a software platform that would allow me to develop and manipulate such elements within a reasonable time frame. After spending considerable time on various internet based software forums, I decided that Macromedia’s Director MX (Macromedia, 2003) application would be a suitable tool. This application originated as one of the earliest multimedia tools but has subsequently been developed to incorporate a substantive object oriented scripting language that can be used to control most modern multimedia file types in real time and interactively with user responses.
Around this time, as outlined in Chapter 1, I had also arrived at the conclusion that two ideal ability constructs for translation into a computer-game-like environment were Processing Speed (Gs) and Working Memory (WM). I decided to first focus on developing a Gs test because I was readily able to arrive at a game-like concept and cover story. Another advantage I considered in conducting research with speed tests was the short amount of time these tests take to complete (i.e., 90 – 180s for standardised tests from, for example, the Wechsler and Woodcock Johnson batteries). I believed this would be a considerable advantage if I needed to conduct extensive piloting of the computer-game-like tests if things went wrong along the way because development time with game-like tasks is usually far greater than for traditional tests. Finally, I was going to need to work within the time limits set out for undergraduate student participation in psychology experiments. The game idea I had was a first person spaceship shooter game based on the Digit Symbol subtest from the Wechsler scales. However, upon arriving at this decision I was confronted with a range of further issues to address before I could proceed to developing the actual computer-game-like test. These issues were addressed in my first study and are described in Paper 1.

Paper 1–Study 1

An important issue I had observed in my previous experience administering computer based tests, with participants of all age groups, was the difficulty many participants had in using the keyboard to respond. Furthermore, with the computer mouse now clearly being the dominant response device for interaction with computers it seemed anachronistic to develop computer-game-like tests which use the keyboard as the primary response device. Indeed, the keyboard promotes a
tendency to look down at the keyboard; a highly undesirable behaviour in tests requiring rapid responses to stimuli presented on the computer screen. For example, Stephens (2001) conducted a case study (i.e., single participant) using eye gaze recordings while a coding task much like Digit Symbol was performed. This coding task used a keyboard response format and the eye gaze data allowed fixation times on different parts of the task to be separately measured. Results suggested that approximately one third of the response time for each item was spent looking at the keyboard.

This may not be a significant problem in a more traditional computerised speed test but this is clearly a problem within a computer-game-like scenario where other important visual information is being presented on the screen (e.g., scoring and timing feedback). Furthermore, I was also looking ahead to the development of a WM task that would require almost constant visual attention to two onscreen tasks simultaneously.

As stated in Paper 1 previous researchers using computerised coding tasks had opted for keyboard response methods. However, I identified one study in which researchers had successfully created speeded search tasks using a mouse response method (Parks, Bartlett, Wickham, & Myors, 2001). Parks et al. found a strong factor underlying paper-and-pencil and mouse response processing speed tasks and a second response method factor which accounted for only a small amount of variance in comparison to the first factor. Thus, it appeared to be a reasonable proposition to develop a mouse based speeded coding task. However, given the relative novelty of the response method, it was necessary to establish whether a simple coding task using a mouse response method would measure Gs before embarking upon the development of a fully fledged computer-game-like test. The rationale was, that if I developed a coding task incorporating game-like
graphics, sounds and scoring elements, for instance, and it failed to measure Gs, it would be unclear which element, or elements, were responsible for diminishing the construct validity. I thus focused my first study on developing and assessing a mouse based coding task (Symbol Digit; Figure A) which was as close as possible to the original paper-and-pencil Digit Symbol coding task, in terms of its cognitive demands, and assessing the extent to which the task was impacted by the mouse response method. It also seemed important to assess whether the new test would share more variance with reaction time (Gt) measures than with Gs tasks because of the similarly psychomotor nature of the response methods.

Piloting of earlier versions of the tasks described in Paper 1 showed that people had considerable difficulty due to their losing the cursor off of the response grid. This was in line with research I had come across showing that people often spent most of their time, when moving a mouse to a target, in overshooting and subsequently correcting for this (Phillips & Triggs, 2001). I then experimented with constraining the cursor to the response grid and found that this, combined with a larger cursor, resulted in greater approval from my respondents and largely resolved the problem. A further improvement resulted from participant feedback indicating that a sound accompanying correct responses would help them perform the task more easily. For this reason I added a ‘laser’ style sound of very short duration for correct responses; this simple feature resulted in greater approval from respondents. Already, even trying to translate the task in as simple form as possible, I was encountering some of the subtle difficulties involved in test translation to a specific computer format: mouse response.

Results for the first study were better than I had expected with the Symbol Digit test showing strong loadings on the Gs factor and not on the reaction time (Gt) factor (see Paper 1 for details). However, there were two issues that stood out
to me as important. First, there was a far greater rate of errors for the computerised tests than the paper-and-pencil tests. Second, the measure of mouse speed I tried to develop (Digit Digit) appeared to be as strong an indicator of Gs as was Symbol Digit. Upon reflection, although I had tried to design the test to measure very simple reactive mouse speed, it seemed obvious that the task still involved much of the visual scanning presumably involved in Gs tasks. Thus, in the subsequent two studies, reported in Paper 2, I aimed to address these issues in addition to developing a more game-like version of Symbol Digit.

Paper 2–Study 1

My conclusion after Paper 1 was that Symbol Digit, using the mouse response method I had developed, had acceptable psychometric properties. The next step was to make the test look like a computer game. The overarching goal for Paper 2 was to use the format of Symbol Digit as a template for a simple ‘first person shooter’ style computer game. However, I decided to introduce game-like
elements in two steps for the same reason I had created Symbol Digit as a first step; in the event of failure I wanted to be able to narrow down the potential reasons for the failure of the test. I thus chose to conduct this smaller scale pilot study first, only incorporating the basic graphical elements of the final game design.

I first embedded the coding task within a spaceship cockpit view with the response grid now sitting on the control panel of the spaceship. Looking out the window, the player would see enemy spaceships instead of symbols (see Figure B). At the bottom of the screen, there were nine spaceships each lined up with a number from one to nine. To destroy any enemy spaceship the player had to fire a laser using the number that matched the spaceship as shown. As a spaceship was destroyed another would arrive. This provided a task (Space Code) that was essentially the same as Symbol Digit but which used more complex graphical stimuli. The space ships used were high resolution full colour graphics. Initial piloting indicated some problems that had also been of some concern in the study reported in Paper 1.

First, the mouse response method encouraged a rapid, error prone response strategy in a small number of individuals; these individuals were often those who reported higher levels of recreational game use. The issue in Symbol Digit was that there was no negative consequence for this, thus making it a sensible strategic option. I opted to try and discourage this response style by using a positive outcome for correct responses and a negative outcome for incorrect responses. A correct response resulted in a ‘laser’ sound and the ship exploding
Figure B. *Screenshot from Space Code showing a practice item.*

(via an animated explosion and accompanying sound) while an incorrect response resulted in a ‘banging’ sound and the screen flashing to indicate to the participant that they had been hit. Piloting suggested that this had having the desired effect. I also incorporated a response method practice session into the game, prior to beginning the game, which I hoped would also serve as a measure of response speed (Mouse-Grid Speed) that was relatively independent of other cognitive aspects of the task.

Second, another problem that some players reported was having trouble keeping track of the cursor. This seemed more pronounced in Space Code than in Symbol Digit, and this was likely due to the darker colours and stronger contrasts on the screen, or to the greater complexity of the screen layout and graphics. I addressed this issue by programming the response grid squares to change to a much brighter colour as the cursor rolled over them so as to provide a clearer interactive indication of where the cursor was. Players found this much easier to
use and they also reported being able to see the cursor’s location with their peripheral vision while looking elsewhere on the screen.

The final consideration in the design of this study was the battery of ability tests to include. On the basis of results from Paper 1, it was deemed appropriate to include Gs tests, as opposed to Gt tests. The addition of strong and complex visual elements in Space Code also suggested that measures of visual abilities (Gv) might also be appropriate to include. Because this study was intended to be a smaller scale pilot study before developing the full game, a small battery of tests measuring Gs and Gv was used. Results from this study suggested that although Space Code did appear to measure Gs, it also measured aspects of Gv. The original Symbol Digit test once again loaded only on a Gs factor, again suggesting that the response method was appropriate for a Gs test. Of additional concern, once again, was the observation that the elements designed to reduce rapid error prone responding strategies had failed to work as well as they did in piloting. Further details are provided below in discussing Study 2.

_Paper 2–Study 2_

For Study 2, I aimed to make Space Code as game-like as possible (Space Code Game) although constrained by the need to keep the principal cognitive demands of the task the same as for the previous versions and, indeed, the traditional coding task. The additional elements added included a scoring system which I was also able to utilize as an additional form of negative feedback for error-prone response strategies (see above). I developed a scoring system that descended visually (i.e., the score scrolling downwards in value) when an error was made and I programmed the response grid to be deactivated for approximately half a second while this occurred. This meant that players could not quickly respond if many
Figure C. Screenshot of Space Code Game showing item in test with score and bonus point stars in upper left and time remaining in upper right. Number to the left of response grid indicates current points available for destroying ships.

errors were made and they would also see their score going backwards. Piloting suggested that the deactivation of the response grid in particular was a much stronger deterrent to error-prone responding than the method used in Study 1. Although not reported in the paper, I found suggestive evidence that error rates had been reduced by these alterations.

Error rates were calculated by dividing the number of errors by the total number of responses (both correct and incorrect). The mean error rate for Space Code Game \( M = .042, SD = .034 \) was slightly lower than for Space Code \( M = .044, SD = .060 \) although a Mann-Whitney Test indicated that this difference was not statistically significant, \( Z = 1.30, p = .192 \). However, the maximum error rate for Space Code Game (.13) was substantially lower than for Space Code (.33) suggesting that the manipulation may have effectively deterred the small number
of participants who may have otherwise adopted an extremely rapid but error prone strategy.

At this point it also seemed important to establish some form of reliability for Space Code Game. Thus, in order to remove the effects of memory for a proposed second administration of the game I developed an alternate form with the same space ships associated with different numbers. However, initial piloting of this version indicated that players were severely distracted by this, with most players making considerable numbers of errors due to responding according to the previously used code. I was unable to conduct the second administration at a later date due to restrictions in participant availability so I would have to work within these constraints as best as possible. I concluded that such a strong interference effect was probably worse for construct measurement validity than any beneficial effect of recent memory in a second administration of the task. This also suggested to me that visual associational memory could impact Space Code Game performance which lead to the inclusion of tests for this construct in the second study.

As in the previous studies, I was interested in further refining the separate measure of response speed (referred to as Mouse-Grid Speed in Paper 2–Study 1). My goal was to reduce cognitive demands of every possible type in an attempt to measure only the speed of moving the cursor around the grid and clicking the mouse. However, each successive version developed so far seemed to also measure Gs to a reasonable extent. The test used in Paper 2–Study 1 (Mouse-Grid Speed) worked well apart from some players having trouble remembering the pattern, or being unclear if their responses had registered, or both of them. To address this I developed a new version in which the pattern was practiced and
thus open to memorization, and each successive target number also lit up a brighter colour so successful performance did not rely only on memory.

The battery of tests used in Study 2 was larger than in Study 1 and included a greater variety of Gs tests. Results of this study suggested that Space Code Game was reliable across two administrations within the same session and that it showed evidence of construct and discriminant validity. The newly developed measure of response speed (Guided Mouse-Grid Speed) also loaded on a Gs factor. It therefore appears that there may be something inherently similar in rapidly using a cursor and performance on Gs tasks. With Space Code (in the form of Space Code Game) developed to this point I began to focus on the development of a new computer-game-like test to measure WM as well and planning studies to assess the construct validity and external validity of both tests. These studies are described next.

*Paper 3–Study 1*

At earlier points in time I had considered a range of different test types that could potentially be instantiated within a computer game format. An initial idea was to have one overall cover story for a game and incorporate different sub tasks within different sections of the game. I thus set about creating an additional task that could be incorporated into the Space Code environment which could potentially measure WM. As WM tasks often use competing task formats it also seemed appropriate to develop a simultaneous task with the coding task performed in Space Code. This also seemed consistent with the familiar situation in computer games where players are required to attend to multiple aspects of a game at the same time.
While contemplating how I might design such a task, I recalled games where one was required to move in and out of different sectors of a space war environment making use of memory for what was in these sectors. At this time I was also programming a battery of computerised WM tasks for a fellow postgraduate student. One of the tasks I was programming, the Dot Matrix task, seemed to me quite similar to the sector grids used in the games I had played. This then formed the basis for the original idea for Space Matrix.

Space Matrix (see Figures D and E) underwent a great deal of piloting in numerous versions before arriving at the final version. I tried the task with the sector grid appearing at intervals in the middle of the screen but this broke up the flow of the game. I tried with the sector grid being displayed to the right of the response grid all the time but this evoked different strategies in players that benefited some but not others very obviously. I then opted to display the grid in this position but for a brief period of time so as to make players rely on apprehension and memory maintenance. I found I then needed to add a sound to indicate the display onset of a new grid location. This further lead to the decision to remove the music and some of the other sounds used in Space Code because the new task was becoming distractingly busy. I also had to refine the length of exposure, numbers of locations to be recalled, and incorporate a bonus scoring system for this aspect of the game so as to optimize performance variance and motivation. The variations created and piloted are too numerous to detail but they indicate a sample of the myriad issues that will need to be considered if computer-game-like assessment tools become more complex than those described in this
Figure D. *Screenshot from Space Matrix showing frame with sector grid being displayed to the right of the response grid in cockpit.*

NOTE: This figure is included on page 46 in the print copy of the thesis held in the University of Adelaide Library.

Figure E. *Screenshot of Space Matrix showing frame with player prompted to recall and enter sectors that have been operated in previously to gain bonus points.*

NOTE: This figure is included on page 46 in the print copy of the thesis held in the University of Adelaide Library.
thesis. As well as assessing the construct validity of the two computer-game-like tests I was also interested in assessing the relationships between game performance and external variables such as gender, handedness and previous computer game experience. I had also assessed these in the earlier studies, but had not reported these in the papers due to word limits. The results relating to these measures were very similar to the results obtained in the previous studies, with gender and game experience showing relationships with each other as well as game and ability measures. These issues will be discussed more fully in the Conclusion of the thesis.

**Paper 3–Study 2**

While I was beginning data collection for Paper 3–Study 1, I also began developing and piloting a version of Space Matrix that was calibrated, in terms of difficulty, for younger children. I was working on this in preparation for a school based study to be completed after Study 1. At this time I was approached (via my PhD supervisor) by a graduate student who was interested in the games I was developing. I explained my plans for a school based study using Space Code and Space Matrix, Gs and WM marker tests, and school grades as external measures. A series of discussions resulted in my co-supervision of the student, Ertimis Eshkevari, and a portion of this research was used for the purposes of her Honours Psychology research project. The student completed the data collection for the project while I was responsible for the development of all tests and overall design of the study and the data has been extensively reanalysed within the present thesis.

The size of the test battery for this study was limited by the availability of high school students for approximately 50 minutes only, including any time taken to
seat them at the computer terminals used. Students were tested in small groups by the graduate student. As noted in the paper there were concerns about some children adopting a trial and error method to learn the games even after the formal practice. Unfortunately this was difficult to address with children being tested in groups of up to four but this was also probably an inherent failure of the practice session design more broadly, as noted in the paper, and discussed in the Conclusion to the thesis.

Despite these shortcomings, and although Space Code showed less validity in this sample than did Space Matrix, there were a number of notable outcomes. Specifically, the finding that Space Matrix was about as strongly related to school grades as was the Raven’s Progressive Matrices test used in the study is most noteworthy. Observations and discussions with the children also suggested that they found the computer-game-like tests more enjoyable than the traditional tests. However, a discussion of this issue is left for the Conclusion.

Further Notes on Papers

As noted earlier each of the papers presented were predominantly my own work and initiative. My original plan was not to produce a thesis using these papers but I was instead motivated by a desire to publish and receive comment and feedback on the research. I believe it is an important area for research and given the time lags in publication I thought it was beneficial to do this promptly and prior to writing a more conventional thesis. However, when I came to writing the thesis, I began to see that the papers were a rather accurate, and honest, reflection of my thinking and rationale throughout the research program.

The reader will no doubt notice a slightly different approach taken in each paper. Sometimes this is due simply to the terminology used for constructs, while
at other times it is a focus on different aspects of the research, or different variants of factor analytic methodology. These variations were due to my own changes in thinking over the course of time but also due to suggestions from reviewers of the papers and made to aid their reading or to elaborate on aspects of the research that they thought readers may be further interested in. I chose the analyses that best fitted the particular datasets at the time and I aimed to provide more comprehensive analyses in the final two studies. An explication of all the variations, and reasons for them, arising from reviewer comments and subsequent revisions is, in my opinion, likely to distract from the overall flow of the work. I trust that the readers, having been through the process themselves on numerous occasions, are suitably content to hold these issues imaginatively in the background whilst reading the final versions of these papers within the broader context of the thesis as set out in the supporting chapters.

To further aid the reader of this thesis, and as part of the examinable content, I have attached a CD-ROM containing each of the tests used within the thesis. These are self executing stand alone applications and are located in folders specific to the three papers and labeled as referenced in each paper. A read me file provides additional supporting information for each application.
A Speeded Coding Task Using a Computer-Based Mouse Response

Jason McPherson and Nicholas R. Burns (2005)

School of Psychology, University of Adelaide

Behavior Research Methods, 37 (3), 538-544

Statement of Contributions

Jason McPherson (Candidate)

I was responsible for the conception and primary authorship of the paper. I was responsible for the development and programming of software based assessment tools and collection of all data. I conducted the statistical analyses independently with advice from the co-author. I was corresponding author and primarily responsible for responses to reviewers and revisions to the paper.

Signed……………………………………..Date……………………………………

Nicholas R. Burns (Co-author)

I was the supervisor (advisor) for the research programme that lead to this publication. In terms of conceptualisation of the programme, there was extensive and ongoing collaboration between Mr. McPherson and me in developing the direction of the research. The realization of the program, specifically, development of the computer-game-like and other computerized tests, data collection, and analyses, were the work of Mr. McPherson. I had an advisory role with respect to selection of the test battery used and on the direction and specifics of the data analyses.
Mr. McPherson was responsible for writing this paper; my role was to comment on drafts, make suggestions on the presentation of material in the paper, and to provide editorial input. I also provided advice on responding to comments by the journal reviewers and editor. I hereby give my permission for this paper to be incorporated in Mr. McPherson’s submission for the degree of PhD in the University of Adelaide.

Signed…………………………………….Date…………………………………….
Abstract

This study assessed whether a speeded coding task that used a computer based mouse response (CBMR) format was a measure of General Processing Speed (Gs). By analyzing the task within a network of tasks representing both traditional Gs tests and reaction time tasks it was shown that a CBMR test can be used to measure the same construct as traditional paper-and-pencil (PP) tests and that this response format does not introduce variance associated with psychomotor performance. Differences between PP and CBMR formats were observed and it is argued that these may provide information on individual differences in performance not available from traditional coding tests.
Introduction

Research articles primarily concerned with coding tasks, or substitution tasks as they are also known, usually begin with some reference to the longstanding and widespread use of such tasks within psychology. These statements appear to be justified. Coding tasks are used in research investigating cognitive abilities underlying psychometric intelligence (Ackerman, Beier, & Boyle, 2002; N.R. Burns & Nettelbeck, 2003; Gignac & Vernon, 2003), aging processes (MacDonald, Hultsch, Strauss, & Dixon, 2003; Salthouse, Letz, & Hooisma, 1994), psychopharmacological effects (Hallam, Olver, McGrath, & Norman, 2003; Mattila, Aranko, Mattila, & Paakari, 1994), health related cognitive impairment (I.J. Deary, Sommerfield, McAulay, & Frier, 2003) and attention (Bate, Mathias, & Crawford, 2001; Pogge, Stokes, & Harvey, 1994). These tasks, the most well known being Digit Symbol from the Wechsler intelligence scales, have their roots in the coding subtest of the Army Beta battery developed around 1917, which was in turn based on an earlier test devised by Otis (Matarazzo, 1972).

Digit Symbol is a paper-and-pencil coding task requiring the association of symbols with digits by reference to a code table. Performance on this task is measured as the number of correctly associated symbols placed beneath a series of digits in a set amount of time, typically 90 – 120 s. Digit Symbol consists of a series of homogenous items of trivial difficulty administered under strict time constraints. These features are common to tests loading on ‘General Processing Speed’ (Gs) within contemporary factor analytic theories of intelligence. Such tests are commonly referred to as ‘speed’ tests.

There has been considerable research investigating the relationship between paper-and-pencil (PP) and computer based (CB) presentations of speed tests (Mead & Drasgow, 1994). The most notable finding is that the cross modal (PP
with CB) correlations for speed tests tend to be lower than for power tests (in which items of varying difficulty are administered with minimal time constraints). Mead and Drasgow estimated a cross modal correlation of .72 for speed tests and .97 for timed power tests. The reason for the lower correlation for speed tests, according to the authors, is that perceptual and motor processes contribute more variance in speed tests than in power tests and that even minor alterations to the presentation and response formats will affect these processes. This is supported by research demonstrating the sensitivity of speeded tests to even slight modifications to response formats (Wegner & Ree, 1985).

In the past, researchers who have developed computerized coding tasks have opted for a keyboard response format (Baker, Letz, & Fidler, 1985; Laux & Lane, 1985; Mattila et al., 1994; Salthouse et al., 1994). These CB tests have achieved a number of stated goals such as allowing the automation of scoring, dynamic manipulation of task components and being sensitive to the effects of drugs (Mattila et al., 1994) and aging (Salthouse et al., 1994). However, it can be argued that using a keyboard response format is discrepant with the way that most people interact with modern computers. Indeed, research investigating eye gaze movements during a CB coding task, found that approximately one third of response time was associated with looking at the keyboard (Stephens, 2001). It would thus seem that although some people may find the keyboard a comfortable way to respond, those who can touch type for instance, many others may find the keyboard an awkward way to respond.

It is our view that the computer mouse provides a more natural and ecologically valid response format than the keyboard. First, because it is the most common device for interacting with a computer used by the general public. Second, because it allows a person to keep their eyes fixed on the screen regardless of
their level of computer experience. Previously, many experimental tasks were programmed as, for example, DOS based applications, and were therefore most amenable to keyboard response formats. However, the general public have long ago abandoned such operating systems in favor of those with a graphic user interface. Just as handwriting was the natural response format for PP tests, we believe the mouse is the natural response format for CB tests.

The general aim of the present study was to develop a speeded coding test utilizing a computer based mouse response format (CBMR). Despite the evidence for mode of presentation effects in CB tests, we believe that the important question is whether a CB coding test can provide a measure of the underlying psychological construct of interest (i.e., Gs). We assume that any test using a CBMR format will have unique variance associated with mouse use just as PP tests have unique variance associated with the graphomotor elements of such tests (Laux & Lane, 1985). However, we do not view this source of variance unfavourably, because, with the increasing use of computers, the inclusion of this variance may potentially increase the ecological validity of Gs tests for some purposes (e.g. personnel selection).

Given the identification of Digit Symbol as a strong indicator of Gs within contemporary theories of intelligence (McGrew & Flanagan, 1998), it is envisaged that a CBMR coding task should also measure this latent construct. However, even if a CBMR test shares considerable variance with traditional Gs tests, the nature of a mouse response format may potentially introduce variance due to psychomotor abilities, such as those commonly measured by reaction time tasks. For this reason, the present study assessed the relationship of a CBMR coding task with both traditional tests of Gs and with paradigmatic reaction time tasks using both exploratory and confirmatory factor analytic procedures as well as
simple correlational analysis. Additionally, the relationships between response formats in PP and CB coding tasks were explored using subtask presentations for each modality.

Method

Participants

There were $N = 49$ participants who were predominantly Level I Psychology students from the University of Adelaide who received course credit for their participation. The mean age was 21.2 years ($SD = 4.32$), 37 were female ($M = 21.4$ years, $SD = 4.5$) and 12 were male ($M = 20.7$ years, $SD = 3.8$).

Materials

Paper-and-pencil tests are denoted by the suffix PP. Computer based tests using a mouse response format are denoted CBMR.

Tests of Gs.

Digit Symbol (DS-PP). This test from the Wechsler Adult Intelligence Scales (WAIS-III) requires the filling in of blank cells according to a key provided at the top of each test sheet. The standard two minute time limit was used.

Visual Matching (VM-PP). This test, taken from the Woodcock-Johnson Psycho-Educational Battery – Revised (Woodcock & Johnson, 1989), requires searching for and circling matching pairs of numbers. The standard three minute time limit was used.

Customised Tests

Symbol Symbol (SS-PP). This paper-and-pencil test was designed to be as similar as possible to the DS-PP test but with the coding element of the task removed. The test is structured in the same way as DS-PP except symbols are
placed in the upper squares and participants are simply required to copy these as quickly as possible. A time limit of one minute was used to minimise fatigue because piloting indicated heavier physical demands for this test than other PP tests. The series of items matched the correct series of responses required in DS-PP.

**Symbol Digit (SD-CBMR).** The layout for this computerised test is shown in Figure 1. Each item consists of a symbol presented in the center of the screen situated within a white box with a black border (5.4 cm x 5.2 cm). A correct response requires participants to left mouse click on the number that corresponds to the symbol in the code table (22.1 cm x 4.7 cm) at the bottom right hand corner of the screen. Participants respond by left clicking the mouse on a 3 x 3 numerical grid (6.8 cm x 6.2 cm) in the bottom right hand corner of the screen. To make the use of this number grid easier, each number’s background turns red when the cursor is placed over it and the cursor is restricted such that it cannot be moved outside of the grid area. When a correct response is made a ‘laser’ style sound is heard and the next item appears in the center of the screen. When an incorrect response is made the computer’s default beep sound is heard and the item remains until the correct response is made. The background color of the screen was blue. Seven practice items are provided before the test begins. Participants are instructed to complete as many items as they can in two minutes. The sequence of responses required are the same as for DS-PP.
Digit Digit (DD-CBMR). This test was designed to be as similar as possible to SD-CBMR, above, but with the coding element removed. The layout of this computerised test is shown in Figure 2. Each item consists of a single digit (between 1 and 9) in Arial Black font situated within a white box with a black border (5.4 cm x 5.2 cm). Participants respond by left clicking the mouse on the corresponding number using the same numerical grid as in SD-CBMR. All other aspects of the response format were also the same as for the SD-CBMR. Seven practice items are provided before the test begins. Participants are instructed to complete as many items as they can in two minutes. The sequence of items are the same as for DS-PP.
Both computerised tests were run on a Pentium 4 class computer (2.4 MHz) with a standard Microsoft mouse. Display was via a 19 inch monitor with a vertical refresh rate of 75 Hz and a screen resolution of 1280 x 1024 pixels.

**Reaction Time Tests.**

**Simple Reaction Time (SRT).** Based on the task described by Jensen (1982). The display and response panel was as described by Jensen and Munro (1979) but scaled down so that the surface of the panel measured 23 cm wide and 15 cm deep; the panel was angled at 30 degrees to the surface of the table. Participants were required to hold down a single button and then release it as quickly as possible after the light above it came on. There were 10 practice trials and 60 experimental trials. The single dependent measure taken from this task was the median time to respond after the onset of the stimulus.

**Odd Man Out Reaction Time.** Based on the task described by Frearson (Frearson, Barrett, & Eysenck, 1988; Frearson & Eysenck, 1986) and using the
same apparatus as for SRT, above. Each stimulus consisted of three of a possible eight lights being lit. One light was the ‘odd man out’; that is, two lights were closer together, with one further away from the pair. The participant responded by pressing the button below the light. There were 20 practice trials and 60 experimental trials. The task returns two dependent variables; median decision time (OMO-DT), measured from the onset of the stimulus array until the participant lifts the finger from the home button; and median movement time (OMO-MT), from release of the home button until the target button is pressed.

Procedure

Participants were tested individually in a quiet room and asked to concentrate as best as they could for each individual test. It was further emphasised that it was important for them to do their best on each test so that an accurate comparison could be made across their performances on all tasks. Test were administered to all participants in the following order: Symbol Symbol (SS-PP); Visual Matching (VM-PP); Digit Symbol (DS-PP); Digit Digit (DD-CBMR); Symbol Digit (SD-CBMR); Simple Reaction Time (SRT); Odd Man Out Reaction Time (OMO-DT and OMO-MT).

Results

Descriptive Statistics

Descriptive statistics for all tests are presented in Table 1. There were no obvious ceiling or floor effects evident in any of the test measures. Considering only the coding tasks, the mean DS-PP score was slightly higher than the mean SD-CBMR score ($t(48) = 1.96, p = .056, d = .57$) However, SD-CBMR scores had a larger range than DS-PP scores, the SD-CBMR test having a lower minimum, higher maximum and a correspondingly higher standard deviation. Inspection of
Table 1

*Descriptive statistics for Reaction Time (ms), Paper-and-pencil and Computer Based Mouse Response Tests (number of correct items).*

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>SD</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digit Symbol (DS-PP)</td>
<td>89.8</td>
<td>12.5</td>
<td>67</td>
<td>124</td>
</tr>
<tr>
<td>Visual Matching (VM-PP)</td>
<td>56.0</td>
<td>6.1</td>
<td>42</td>
<td>67</td>
</tr>
<tr>
<td>Symbol Symbol (SS-PP)</td>
<td>95.5</td>
<td>13.2</td>
<td>68</td>
<td>133</td>
</tr>
<tr>
<td>Symbol Digit (SD-CBMR)</td>
<td>86.4</td>
<td>15.9</td>
<td>61</td>
<td>128</td>
</tr>
<tr>
<td>Digit Digit (DD-CBMR)</td>
<td>55.9</td>
<td>6.1</td>
<td>42</td>
<td>67</td>
</tr>
<tr>
<td>Simple Reaction Time (SRT)</td>
<td>218.5</td>
<td>25.2</td>
<td>170</td>
<td>299</td>
</tr>
<tr>
<td>Odd Man Out Decision Time (OMO-DT)</td>
<td>531.6</td>
<td>116.3</td>
<td>351</td>
<td>929</td>
</tr>
<tr>
<td>Odd Man Out Movement Time (OMO-MT)</td>
<td>206.8</td>
<td>42.1</td>
<td>130</td>
<td>320</td>
</tr>
</tbody>
</table>

The distributions for each test suggested that SD-CBMR test scores were more positively skewed than DS-PP scores, with skewness coefficients of .63 and .27, respectively.

*Correlations*

Correlations between test scores and age indicated age related effects that were not consistent across tests. For example, the correlations with age for OMO-DT and SD-CBMR were .61 and .06, respectively. For this reason test score variables were age residualized for subsequent analyses. The correlations between all tests are presented in Table 2.
**Table 2**

**Intercorrelations for Reaction Time, Paper-and-pencil and Computer Based Mouse Response Tests.**

<table>
<thead>
<tr>
<th>Variable</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Digit Symbol (DS-PP)</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Visual Matching (VM-PP)</td>
<td></td>
<td>.55**</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Symbol Symbol (SS-PP)</td>
<td></td>
<td>.45**</td>
<td>.42**</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Symbol Digit (SD-CBMR)</td>
<td></td>
<td>.67**</td>
<td>.50**</td>
<td>.12</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Digit Digit (DD-CBMR)</td>
<td></td>
<td>.51**</td>
<td>.35*</td>
<td>.26</td>
<td>.56**</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Simple Reaction Time (SRT)</td>
<td></td>
<td></td>
<td>.01</td>
<td>.01</td>
<td>- .06</td>
<td>-.01</td>
<td>-.05</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Odd Man Out Decision Time (OMO-DT)</td>
<td></td>
<td>-.13</td>
<td>-.13</td>
<td>.08</td>
<td>-.16</td>
<td>-.14</td>
<td>.30*</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Odd Man Out Movement Time (OMO-MT)</td>
<td></td>
<td>-.06</td>
<td>-.13</td>
<td>-.18</td>
<td>-.20</td>
<td>-.17</td>
<td>.42**</td>
<td>.28</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>9. Symbol Digit Error Rate (SDER-CBMR)</td>
<td></td>
<td>-.17</td>
<td>-.09</td>
<td>-.05</td>
<td>-.12</td>
<td>.07</td>
<td>-.12</td>
<td>-.07</td>
<td>.06</td>
<td>1</td>
</tr>
<tr>
<td>10. Digit Digit Error Rate (DDER-CBMR)</td>
<td></td>
<td>.36*</td>
<td>.25</td>
<td>.21</td>
<td>.37**</td>
<td>.33*</td>
<td>-.36*</td>
<td>-.26</td>
<td>-.34*</td>
<td>.48**</td>
</tr>
</tbody>
</table>

* p < .05, ** p < .01
The two Gs marker tests, VM-PP and DS-PP, were correlated .55 and the largest correlation in the matrix is between DS-PP and SD-CBMR. The correlations between the two Gs marker tests and the new SD-CBMR test indicate substantial common variance between these three tests. The reaction time tasks appear to share common variance indicated by the pattern of small-to-moderate significant intercorrelations of these tasks and their near-zero correlations with the PP and CBMR tests.

**Exploratory Factor Analysis (EFA)**

The validity of SD-CBMR as a measure of Gs was assessed using EFA. This was done prior to any Confirmatory Factor Analysis (CFA) to allow any inherent structure in the data to be assessed independently of theoretical expectations. The SD-CBMR test together with the three reaction time tasks and the two Gs marker tests were entered into an initial Principal Components Analysis (PCA). There were two components with eigenvalues larger than one (2.28 and 1.57) accounting for 38.0% and 26.2% of the variance, respectively. Inspection of the scree plot also indicated two dominant components. Using the eigenvalue and scree criteria as an initial guide, and in an attempt to avoid an under estimation of underlying factors (Fabrigar, Wegener, MacCallum, & Strahan, 1999), it was decided to begin further analysis with the extraction of three factors.

Because we were primarily interested in the latent structure of the data and given that skew and kurtosis for all of the variables were below one, maximum likelihood extraction was used (Fabrigar et al., 1999). An attempt to extract three factors resulted in a Heywood case (estimated communality for a variable larger than one) and a third factor almost identical to the first factor. Extracting two factors resulted in a nonsignificant Chi-Square ($\chi^2 = 1.31, df = 4, p = .860$) indicating an acceptable fit. Direct Oblimin rotation resulted in two clearly defined
factors \( r = -.19 \), each saliently loaded by three tests; largely identical results were obtained with a Varimax rotation. A one factor solution indicated small loadings for the reaction time tests and the Chi-Square \( \chi^2 = 15.92, df = 9, p = .068 \) suggested a relatively poorer fit. The two factor Varimax solution was thus preferred on grounds of fit, simplicity and interpretability. The pattern of factor loadings for the final solution is presented in Table 3.

Factor 1 is saliently loaded by the two Gs marker tests and SD-CBMR with very small loadings from the three reaction time tests. This factor is thus interpreted as Gs. Factor 2 is saliently loaded by the three reaction time variables with very small loadings from the other tests. This factor is interpreted as a reaction speed factor. This pattern of factor loadings supports the validity of the SD-CBMR test as a Gs measure.

**Table 3**

*Factor loadings of test variables for Maximum Likelihood Extraction with Varimax Rotation*

<table>
<thead>
<tr>
<th>Variable</th>
<th>Factor 1</th>
<th>Factor 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digit Symbol (DS-PP)</td>
<td>.857</td>
<td>-.012</td>
</tr>
<tr>
<td>Visual Matching (VM-PP)</td>
<td>.633</td>
<td>-.077</td>
</tr>
<tr>
<td>Symbol Digit (SD-CBMR)</td>
<td>.783</td>
<td>-138</td>
</tr>
<tr>
<td>Simple Reaction Time (SRT)</td>
<td>.055</td>
<td>.629</td>
</tr>
<tr>
<td>Odd Man Out Decision Time (OMO-DT)</td>
<td>-.141</td>
<td>.439</td>
</tr>
<tr>
<td>Odd Man Movement Time (OMO-MT)</td>
<td>-.088</td>
<td>.670</td>
</tr>
</tbody>
</table>

*Note.* Bold numbers indicate salient loadings
Confirmatory Factor Analysis (CFA)

CFA was undertaken to assess the fit of a measurement model based on the EFA results. First, a two factor model was specified in which the three reaction time measures were allowed to load on one latent variable and the two Gs markers and the SD-CBMR test were allowed to load on the other. The latent variables were specified as independent. The fit of this model to the data was very good ($\chi^2 = 4.26$, $df = 9$, $p = 0.894$, GFI = .972, RMSEA < .001). To assess the potential relationship of reaction time speed with SD-CBMR the initial two factor model was modified by allowing SD-CBMR to load on both latent variables. If SD-CBMR shared any substantial variation with the reaction time tests we would expect an improvement in model fit. This respecified model also fit the data well ($\chi^2 = 3.27$, $df = 8$, $p = .917$, GFI = .979, RMSEA < .001) but a nested model test indicated that the Chi-Square difference was nonsignificant ($\Delta \chi^2 = .99$, $\Delta df = 1$, $p = 0.317$). The estimated path loading between the latent reaction time variable and SD-CBMR was -.14. These results further support the validity of the SD-CBMR as a measure of Gs that is unconfounded with reaction time.

Subtask analyses

The SS-PP and DD-CBMR tests served as subtasks of the DS-PP and SD-CBMR tests, respectively, in which the same presentation and response formats were used but with no coding of stimuli required. It can be seen in Table 2 that these two subtasks are moderately correlated ($r = .26$) but they are more strongly related to their corresponding full task equivalents. The SS-PP test does not appear to have a substantial relationship with the computerized coding task, SD-CBMR ($r = .12$), whereas DD-CBMR seems to share considerable variance with DS-PP ($r = .51$) and, to a lesser extent, VM-PP ($r = .35$).
To compare the mean scores for each subtask we multiplied SS-PP scores by two to adjust for the 1 minute time limit. The adjusted mean SS-PP score (190.1, $SD = 26.4$) was clearly higher than the mean DD-CBMR score ($t(48) = 13.1$, $p < .001$, $d = 3.8$).

**Analysis of Errors**

One further observed difference between the test modalities was the error rate. In all of the PP tests the number of errors made were so few as to make any analysis of individual differences meaningless. However, using the mouse response formats, participants varied more considerably on the numbers of errors made in the DD-CBMR ($M = 4.0$, $SD = 4.8$) and DS-CBMR ($M = 5.3$, $SD = 5.3$). These error scores were converted to error rates to control for speed accuracy trade-off and then age residualized to make these measures consistent with the other test measures. Error rates for each of the CBMR tests were correlated .48 indicating a moderate level of shared variance but a substantial amount of unique variance as well. This is further evident in the differential correlations with the other tests. Error rates for DD-CBMR (DDER-CBMR) had larger correlations with other test measures than error rates for SD-CBMR (SDER-CBMR) had with the same tests.

**Discussion**

This study assessed the validity of a coding task that used a computer based mouse response (CBMR) format as a measure of General Processing Speed (Gs). By analyzing the task within a small network of tasks representing both traditional Gs tests and reaction time tasks it was shown that a CBMR test can be used to measure the same construct as traditional paper-and-pencil (PP) tests and that this response format does not introduce variance associated with psychomotor
performance. These conclusions are supported by correlational and exploratory and confirmatory factor analytic results.

The test developed for the present study, SD-CBMR, had similar correlations with the two paper-and-pencil Gs marker tests as these did with each other. Furthermore, these correlations were possibly attenuated in the largely university based sample. Thus, the correlation between the SD-CBMR and the traditional Gs measures in a more representative sample could potentially be larger than estimated in the present study. Despite this possibility, the intercorrelations between tasks were large enough for a distinct pattern to emerge in EFA.

The EFA solution suggested two common factors underlying the six speed tests. We interpreted the first factor as Gs and the second as a reaction time factor. SD-CBMR was clearly loaded on Gs and not on the reaction time factor and this finding was supported in the subsequent CFA. These results illustrate the advantages of using factor analytic techniques and support Mead and Drasgow’s (1994) appeal for the use of such methods. However, despite the similarities between the PP and CBMR tests it should be noted that a number of differences were observed in the present study.

Test scores from DS-CBMR were slightly lower than for DS-PP which indicates that the CBMR coding task may be more difficult in some way. In parallel with this result, the PP subtask, SS-PP, had an even higher mean relative to the CBMR subtask, DD-CBMR. This suggests that for most people the CBMR format was slower than the traditional PP response format. However, the larger range and skew of scores for the SD-CBMR test suggest this was not the case for all people. Interpretation of these differences is made difficult by the fact that PP and CBMR tests were differentiated by more than the use of mouse or pencil. First, each test modality required differential perceptual scanning. In the traditional DS-PP the
scanning distance from item to code table and back alters as one progresses through the test. Second, the traditional test also permits people to use their previous responses as an additional and potentially more convenient code table proxy. Both of these factors, in addition to the use of the mouse, could have contributed to differences in test means. This is further supported by the intercorrelations of the two subtasks with other tests.

The subtasks for each modality, SS-PP and DD-CBMR, were not highly correlated with one another, suggesting that they do not tap a common speed construct. However, each test had more substantial relationships with other tests. SS-PP was moderately but significantly correlated with DS-PP and VM-PP while it had a negligible relationship with SD-CBMR. It thus seems likely that the relationship between SS-PP and the two paper-and-pencil Gs measures was largely due to the graphomotor and feature encoding requirements in these tests. In contrast, DD-CBMR was more substantially correlated with all of the tests that loaded on Gs, particularly DS-PP and SD-CBMR. This suggests that DD-CBMR required some ability that all the Gs test shared, especially the coding tasks.

Although, DD-CBMR would not seem to require the type of perceptual scanning that coding tasks do, it does require rapid eye movement between the center of the screen and the response grid and the subsequent movement of the cursor to the correct digit. The most obvious suggestion, then, is that these visual requirements of DD-CBMR are related to those required in coding tasks in general. In this sense, the subtask was not entirely successful in that it did not serve as a satisfactory measure of simple response speed. However, the analysis of errors in the CBMR tasks does provide some additional insight into the differences between PP and CBMR formats.
It was clear that the CBMR format evoked higher error rates than the traditional PP formats which are generally found to have negligible error rates. The first conclusion that might be drawn on the error rates is that it is easier to make an error using the CBMR response format. Although the CBMR format evoked slower performances overall, the final action of clicking the mouse is potentially quicker than the drawing of a symbol and may thus be more prone to errors associated with speed accuracy trade-offs. Furthermore, although the error rates in SD-CBMR and DD-CBMR were similar, their intercorrelations with other measures suggest that error rates from DD-CBMR were more substantially related to other speed tests. One possible interpretation of this is that the simpler DD-CBMR is more sensitive to a rapid responding style that may lead to speed accuracy trade-offs.

However, despite all of this, error rates in SD-CBMR were minimally correlated with every measure except the error rate from DD-CBMR. This moderately high correlation of error rates for each CBMR test is reminiscent of so called ‘carelessness’ constructs that have been identified in factor analytic studies of attentional abilities (Carroll, 1993). Thus, although SD-CBMR scores do not appear to confound speed-accuracy trade-offs with overall speed scores, error rates from the test may potentially provide additional information about a person’s response style and attentional abilities.

Although using a CBMR format is unlikely to provide a parallel form of any traditional speed test, the present study suggests that this format can be used in the assessment of speeded ability constructs. However, the differences between traditional PP and CBMR formats should not be overlooked because they may prove useful in some instances and problematic in others. Additionally, before CBMR formats can be more generally accepted, further research will be necessary to assess the external validity of any tests that utilise such a response format.
Presently though, we remain optimistic that the psychological assessment of speeded constructs can cautiously move with the technologies that now surrounds us.
Gs Invaders: Development and Assessment of a Computer-Game-Like Test of Processing Speed

Jason McPherson and Nicholas R. Burns (2007)

School of Psychology, University of Adelaide

Behavior Research Methods, 39 (4), 876-883

Statement of Contributions

Jason McPherson (Candidate)

I was responsible for the conception and primary authorship of the paper. I was responsible for the development and programming of software based assessment tools and collection of all data. I conducted the statistical analyses independently with advice from the co-author. I was corresponding author and primarily responsible for responses to reviewers and revisions to the paper.

Signed…………………………………….Date…………………………………….

Nicholas R. Burns (Co-author)

I was the supervisor (advisor) for the research programme that lead to this publication. In terms of conceptualisation of the programme, there was extensive and ongoing collaboration between Mr. McPherson and me in developing the direction of the research. The realization of the program, specifically, development of the computer-game-like and other computerized tests, data collection, and analyses, were the work of Mr. McPherson. I had an advisory role with respect to selection of the test battery used and on the direction and specifics of the data analyses.
Mr. McPherson was responsible for writing this paper; my role was to comment on drafts, make suggestions on the presentation of material in the paper, and to provide editorial input. I also provided advice on responding to comments by the journal reviewers and editor.

I hereby give my permission for this paper to be incorporated in Mr. McPherson’s submission for the degree of PhD in the University of Adelaide.

Signed…………………………………….Date…………………………………………
Abstract

Computer games potentially offer a useful research tool for psychology but there has been little use made of them in assessing cognitive abilities. Two studies assessing the viability of a computer-game-like test of cognitive processing speed are described. In Study 1 \((N=60)\) a computerized coding task that uses a mouse response method (McPherson & Burns, 2005) was the basis for a simple computer-game-like test. In Study 2 \((N=61)\) dynamic game-like elements were added. Validity was assessed within a factor analytic framework using standardized abilities tests as marker tests. We conclude that computer-game-like tests of processing speed may provide an alternative or supplementary tool for research and assessment. There is clearly potential to develop game-like tests for other cognitive abilities.
Introduction

Psychologists have used computers for at least three decades, with uses ranging from the automation of simple questionnaire items to assistance in taking complex psychological histories (Epstein & Klinkenberg, 2001). Computerization of clinical instruments, personality scales, job attitude scales and various types of cognitive tests has lead to many advantages including ease of administration and standardization, less training for proctors, fewer scoring errors, faster (even immediate) results, fewer disposable materials and, in some cases, fewer opportunities for cheating (Epstein & Klinkenberg, 2001; Mead & Drasgow, 1994). In the area of cognitive abilities assessment, the potential of these advantages has been a motivating force behind large scale projects such as the computerized batteries developed by the US military (Alderton et al., 1997; Peterson et al., 1990; Ree & Carretta, 1998) and by various educational institutions (Bennett, 1998; Mead & Drasgow, 1994). Furthermore, many potential developments in abilities testing have been identified that are largely dependent on tests being computerized (Embretson, 2003). It would thus seem that computerization is certain to play a major role in abilities testing in the 21st century.

Most computerized abilities testing has involved the translation of traditional paper-and-pencil (PP) tests to computer format and it appears that adequately designed computerized tests are potentially as valid as traditional PP tests (Embretson, 2003; Epstein & Klinkenberg, 2001; Mead & Drasgow, 1994; Van de Vijver & Harsveld, 1994). However, given that modern computers are capable of much more dynamic presentation of stimuli than is the case with PP tests, it seems a limited utilisation of a very powerful tool simply to imitate via computer the administration of tests not designed with computerization in mind. Better use of the
distinctive qualities of computer technology may bring additional advantages for both test administrators and testees alike.

One format that does make use of the distinctive features of computerization is the computer game (Donchin, 1995; Porter, 1995; Washburn, 2003). Although some researchers interested in studying cognitive processes have used computer game formats (Berger et al., 2000; Donchin, 1995; Graham et al., 1985; Porter, 1995; Ryan, 1994), there appears to have been less research investigating their use in assessing individual differences in cognitive abilities. Rabbitt, Banerji and Szymanski (1989) studied the relationship between IQ and various measures obtained from a computer game called Space Fortress. Space Fortress was developed to assess the role of practice in skill acquisition (Donchin, 1995) and involves a rather complex set of operations. Rabbitt et al. found that performance on Space Fortress was related to IQ but only after a substantial amount of practice on the game. They argued that practice was necessary to overcome any pre-existing differences in familiarity with computer games. Jones, Dunlap and Bilodeau (1986) assessed the relationship between five arcade style video games and a battery of abilities tests and found that game performance was moderately related to all of the abilities tests they used except verbal ability tests. Similar results were obtained by Jackson III, Vernon and Jackson (1993) who found that a customized computer game requiring dynamic spatial tracking loaded on a factor with the Performance but not the Verbal tests from the Multidimensional Aptitude Battery.

Although these studies suggest a relationship between performance on computer games and cognitive abilities, we believed it would prove more fruitful to design a computer-game-like test from first principles with the intention of measuring specific cognitive abilities. Such systematic use of computer games to
measure psychometric constructs seems to be limited to researchers who have used game-like elements in traditional tests such as Anderson’s (1992) use of alien characters when measuring inspection time in children. It is likely that many other researchers have employed game like elements such as this to engage the interest of children (Berger et al., 2000) but none of these appear to have been further developed or validated as psychometric measures.

The goal of the present study was to assess the construct validity of a simple computerized game-like test. Although the distinction between computer based tasks and computerized games is probably a blurred one (Washburn, 2003), we used game-like graphics and sound, a cumulative scoring and feedback system, and a relevant cover story to provide a more game-like experience for users. We based this game-like test on our simple computerized coding task (McPherson & Burns, 2005) tentatively identified as a measure of Processing Speed (Gs) within the framework of the Cattell-Horn-Carroll (CHC) theory of cognitive abilities (McGrew & Flanagan, 1998). We believe CHC provides an ideal framework for evaluating new tests because many traditional tests have been identified as marker tests for the broad ability domains described within this theory.

CHC theory provides a structural model of human cognitive abilities representing the main points of convergence within modern abilities research (McGrew & Flanagan, 1998). The model incorporates at least two strata but arguably three. The first stratum comprises over 60 abilities that are narrow and which correspond conceptually to Thurstone’s Primary Mental Abilities (1938; 1947). The pattern of inter-correlations among these first-stratum abilities defines about ten broad abilities at the second stratum. These second-stratum abilities include fluid and crystallized intelligence, as first described by Cattell (1943; 1963) along with others more recently described, primarily by Horn and co-workers
(1997). Controversy remains on the interpretation of the pattern of inter-correlations among these broad second-stratum abilities. Many researchers and theorists argue for the existence and primacy of a third-stratum general factor \( (g; \text{ see Carroll, 1993; Jensen, 1998}) \) while others dismiss the importance of \( g \) (e.g., (Stankov, 1998). Deary and Caryl (1997) make the point that whether the broad second-stratum abilities or the putative third-stratum general factor are considered most important depends on the particular research question at hand. Considering individual tests are generally taken to represent the first-stratum the goals of the present study seemed best suited to analysis at the second-stratum.

In summary, we aimed to assess whether the game-like coding task would measure \( G_s \). Considering our use of fairly complex graphics as stimuli, it also seemed appropriate to assess how game performance might also relate to the domain of Visual Processing \( (G_v) \). These broad aims were addressed in two separate studies. In Study 1 our simple computerized coding task (McPherson & Burns, 2005) was altered by substituting more game-like graphics and sound. In Study 2 this modified test was expanded to incorporate some of the more dynamic elements common to computer games. Study 2 also incorporated single marker tests for two other broad abilities to assess discriminant validity. The rationale for conducting two separate studies was to first assess the effect of the more complex game-like visual stimuli before adding more dynamic game elements.

Study 1 – Method

Participants

Full datasets were obtained for \( N = 60 \) participants. Thirty five participants were Level I Psychology students from the University of Adelaide who received course credit for their participation and 25 were students recruited more generally from the
university and paid $20(AU) for their participation. The mean age for the whole sample was 20.6 years ($SD = 2.81$), 32 were female (Mean age = 20.5 years, $SD = 2.27$) and 28 were male (Mean age = 20.8 years, $SD = 3.36$).

**Materials**

**Tests of Gs**

**Digit Symbol.** This test from the Wechsler Adult Intelligence Scale (WAIS-III) requires the filling in of blank cells according to a key provided at the top of each test sheet. The standard two minute time limit was used.

**Visual Matching.** This test, taken from the Woodcock-Johnson III Tests of Cognitive Abilities (WJ-III COG; Woodcock, McGrew, & Mather, 2001) requires searching for and circling matching pairs of numbers. The standard three minute time limit was used.

**Tests of Gv**

**Block Design.** This test from the WAIS-III requires the replication of two-coloured models using blocks. The designs are progressively difficult and scores reflect both accuracy and speed.

**Spatial Relations.** This test is taken from the WJ-III COG (Woodcock et al., 2001). The participant is required to identify the two or three pieces that make up a larger target shape.

**Custom Computer Tests**

**Symbol Digit.** The layout for this computerized test is as described in detail elsewhere (McPherson & Burns, 2005). Each item consists of a symbol presented in the centre of the screen situated within a white box with a black border (5.4 cm x 5.2 cm). Participants respond by left clicking the mouse on a 3 x 3 numerical grid (6.8 cm x 6.2 cm) in the bottom right hand corner of the screen. To make the use
of this number grid easier, each number’s background turns red when the cursor is placed over it and the cursor is restricted such that it cannot be moved outside of the grid area. The correct response is determined according to the code table (22.1 cm x 4.7 cm) at the bottom left hand side of the screen. When a correct response is made a ‘laser’ style sound is heard and the next item appears in the centre of the screen. When an incorrect response is made the computer’s default beep sound is heard and the item remains until the correct response is made. The background color of the screen was blue. Seven practice items are provided before the test begins. Participants were instructed to complete as many items as possible in two minutes. The sequence of responses required is the same as for Digit Symbol.

Space Code. The screen layout for this test consisted of a spaceship cockpit with a numerical response grid of the same dimensions as that used for Symbol Digit and situated in the centre of the screen. However, the number grid is dark blue with white digits. To make the use of this number grid easier, each number’s background turns light blue when the cursor is placed over it and, as with Symbol Digit, the cursor is restricted such that it cannot be moved outside of the grid area. Each item consists of a single spaceship (approximately 5cm x 5cm) appearing in the window view of the cockpit. At the bottom of the cockpit the nine spaceships are each presented with a single digit placed directly above. Participants are instructed by on-screen text that to destroy a ship requires the firing of the number placed above the matching ship at the bottom of the screen. When a correct response is made a laser sound is heard followed by an audible and visible explosion of the spaceship. When an incorrect response is made the screen flashes and a ‘banging’ sound indicates that the participant’s ship has been hit. Seven practice items are provided before the test begins. Participants are
instructed to be as fast and accurate as they can and to destroy as many
spaceships as possible in two minutes. The sequence of responses required is the
same as for Digit Symbol.

**Mouse-Grid Speed.**

This test used the same response grid layout as used in Space Code (see
above). Participants practised clicking two simple patterned sequences of digits
(digits 1-9 in order and 1-3-9-7-5) as quickly as they could for 20 seconds.
Following this they were asked to remember a final sequence (2-6-8-4-5) before
being asked to click this pattern as quickly as possible for 20 seconds. The
number of correct clicks in this final trial served as the dependent measure for this
test. This test was included to provide a simple measure of how quickly
participants could use the response grid independently of the cognitive demands
of the coding operation.

All computerized tests were run on a Pentium 4 class computer (2.4 MHz) with
a Microsoft optical Intellimouse which was brand new at the start of the
experiment. Display was via a 19 inch monitor with a vertical refresh rate of 100
Hz and a screen resolution of 1280 x 1024 pixels.

**Procedure**

Participants were tested individually in a quiet room and asked to concentrate
as best they could for each individual test. It was further emphasized that it was
important for them to do their best on each test so that an accurate comparison
could be made across their performances on all tasks. Tests were administered to
all participants in the following order: Visual Matching; Digit Symbol; Mouse-Grid
Speed; Space Code; Block Design; Spatial Relations; Symbol Digit. Sessions took
approximately 1 hour to complete.
Study 1 - Results

Descriptive Statistics

Descriptive statistics for all tests are presented in Table 1.

Table 1

Descriptive statistics for ability test scores and computer based mouse response test scores (N=60).

<table>
<thead>
<tr>
<th>Measure</th>
<th>Mean</th>
<th>SD</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digit Symbol</td>
<td>89.8</td>
<td>11.8</td>
<td>67</td>
<td>125</td>
</tr>
<tr>
<td>Visual Matching</td>
<td>54.7</td>
<td>6.3</td>
<td>44</td>
<td>69</td>
</tr>
<tr>
<td>Block Design</td>
<td>54.7</td>
<td>8.3</td>
<td>34</td>
<td>68</td>
</tr>
<tr>
<td>Spatial Relations</td>
<td>74.6</td>
<td>3.9</td>
<td>63</td>
<td>80</td>
</tr>
<tr>
<td>Symbol Digit</td>
<td>84.9</td>
<td>14.7</td>
<td>63</td>
<td>132</td>
</tr>
<tr>
<td>Space Code</td>
<td>62.6</td>
<td>9.2</td>
<td>43</td>
<td>83</td>
</tr>
<tr>
<td>Mouse-Grid Speed</td>
<td>56.7</td>
<td>11.2</td>
<td>39</td>
<td>92</td>
</tr>
</tbody>
</table>

Correlations

Correlations between age and test scores indicated a small relationship with Mouse-Grid Speed \((r = -.23, p = .074)\). Correlations between sex and test scores indicated a small but significant relationship with Mouse-Grid Speed \((r = -.28, p = .03)\) and small relationships with Digit Symbol \((r = -.25, p = .057)\) and Space Code \((r = .24, p = .062)\), positive correlations indicate higher scores for males and
Table 2

*Correlation matrix for ability test scores and computer based mouse response test scores (N=60).*

<table>
<thead>
<tr>
<th>Variable</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Digit Symbol</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Visual Matching</td>
<td>.49**</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Block Design</td>
<td>.16</td>
<td>.45**</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Spatial Relations</td>
<td>-.10</td>
<td>.17</td>
<td>.55**</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Symbol Digit</td>
<td>.53**</td>
<td>.41**</td>
<td>.25</td>
<td>.15</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Space Code</td>
<td>.40**</td>
<td>.55**</td>
<td>.53**</td>
<td>.36**</td>
<td>.65**</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>7. Mouse-Grid Speed</td>
<td>.19</td>
<td>.38**</td>
<td>.22</td>
<td>.21</td>
<td>.31</td>
<td>.36**</td>
<td>-</td>
</tr>
</tbody>
</table>

* p < .05, ** p < .01
negative correlations indicate higher scores for females. Inter-correlations between all tests are presented in Table 2. As expected, the two Gs marker tests, Digit Symbol and Visual Matching, were moderately correlated and the two Gv marker tests, Block Design and Spatial Relations, were also moderately correlated.

Considering the computer based mouse response (CBMR) tests it can be seen that Symbol Digit and Space Code share a considerable amount of variance. It can also be seen that Mouse-Grid Speed, which was designed to measure mouse response speed independently of coding speed, shared a similar amount of variance with each of the two CBMR tests. However, the partial correlation between Symbol Digit and Space Code controlling for Mouse-Grid Speed was still .60, \( p < .001 \), suggesting that mouse speed was not the primary determinant of performance in these tasks.

Exploratory Factor Analysis (EFA)

The construct validity of Symbol Digit and Space Code was further assessed using EFA. In addition to Symbol Digit and Space Code, the two Gs markers and two Gv markers were submitted to Principal Axis Factoring (PAF). PAF was used because the primary goal was to identify common factors. Mouse-Grid Speed was not included so as to reduce the possibility of the CBMR tests altering the measurement space too strongly by introducing extra method variance. The number of factors to extract was decided by inspection of the scree plot (Cattell, 1966), parallel roots analysis (Montanelli & Humphreys, 1976), and the theoretical issues relevant to the present study.

Inspection of the scree plot and the parallel roots analysis both indicated a two factor solution. A two factor solution was also deemed appropriate considering the theoretical expectations of separate Gs and Gv factors. The two factor PAF solution was submitted to Promax rotation (Kappa = 4) which resulted in two
Table 3

*Structure Matrix for Principal Axis Factoring of ability test scores and computer based mouse response test scores with Promax Rotation (Kappa = 4)*

<table>
<thead>
<tr>
<th>Variable</th>
<th>Processing Speed (Gs)</th>
<th>Visual-Spatial (Gv)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digit Symbol</td>
<td>.724</td>
<td>.061</td>
</tr>
<tr>
<td>Visual Matching</td>
<td>.668</td>
<td>.415</td>
</tr>
<tr>
<td>Block Design</td>
<td>.433</td>
<td>.769</td>
</tr>
<tr>
<td>Spatial Relations</td>
<td>.148</td>
<td>.725</td>
</tr>
<tr>
<td>Symbol Digit</td>
<td>.724</td>
<td>.316</td>
</tr>
<tr>
<td>Space Code</td>
<td>.772</td>
<td>.637</td>
</tr>
</tbody>
</table>

*Note.* Bold numbers indicate loadings > .60; Factors are correlated .27 and named on the basis of marker tests.

weakly correlated factors \((r = .27)\) each with at least three loadings larger than .5.

This solution accounted for 57.7% of the variance. The structure matrix of factor loadings is presented in Table 3. The Pattern Matrix indicated almost identical results.

Using the marker tests as a primary guide it can be seen that the first factor is strongly loaded by both Gs marker tests and thus supports the interpretation of this factor as representative of this construct. Similarly, the second factor is strongly loaded by the two Gv tests and is correspondingly interpreted as a Gv factor. Looking at the loadings for the CBMR tests, Symbol Digit has a salient loading only on the Gs factor while Space Code has substantial loadings on both.

*Correlations with Composite Measures*

To further assess the relationship between the CBMR tests and the underlying ability constructs, composite Gs and Gv scores were derived using only the
traditional marker tests. Gs composites were derived by computing average $z$ scores for Digit Symbol and Visual Matching. Gv composites were derived computing average $z$ scores for Block Design and Spatial Relations. These results closely match those obtained through EFA. Mouse-Grid Speed was moderately correlated with Gs scores ($r = .33, p = .009$) but not with Gv scores ($r = .14, p = .289$). Similarly, Symbol Digit was more strongly related to Gs scores ($r = .55, p < .001$) than to Gv scores ($r = .22, p = .085$). However, Space Code was related to both Gs ($r = .55, p < .001$) and Gv scores ($r = .50, p < .001$). It is notable that Symbol Digit and Space Code showed correlations with the Gs scores comparable to the correlation between the two traditional measures comprising the composite ($r = .49, p < .001$). Space Code also showed a correlation with Gv scores comparable to the correlation between the two Gv markers ($r = .55, p < .001$).

Study 1 – Discussion

Results from Study 1 support previous research indicating that the Symbol Digit test using a computer-based mouse response method can be interpreted as a measure of Gs (McPherson & Burns, 2005). However, although the Space Code test loaded on the same Gs factor it also had a loading of a similar magnitude on the factor loaded by the Gv marker tests. This raises the possibility that this test may be a mixed measure of Gs and Gv. However, considering that the demands of the task seem qualitatively different from the Gv marker tasks, and given the small sample size of the present study, we were cautious about this interpretation.

With this in mind we decided to proceed with a second study incorporating as many dynamic game-like elements as possible into Space Code. We also increased the number of marker tests used for each construct, included single measures of two other potentially relevant constructs, and assessed the reliability
of the new game-like version. Additionally, we endeavoured to address the issue of a certain responding strategy that was noted as being potentially problematic in the first study. For the Space Code test some participants were observed to respond very rapidly without concern for accuracy within the Space Code test. It was therefore deemed important to incorporate into the Space Code protocol some form of disincentive towards this strategy of responding.

Study 2 - Method

Participants

Full datasets were obtained for \( N = 61 \) participants. All participants were Level I Psychology students from the University of Adelaide who received course credit for their participation. The mean age for the whole sample was 20.0 years \( (SD = 3.56) \), 37 were female (Mean age = 20.3 years, \( SD = 4.25 \)) and 24 were male (Mean age = 19.7 years, \( SD = 2.14 \)).

Materials

Tests of Gs

Digit Symbol. Described in Study 1.


Decision Speed. This test from the WJ-III COG requires searching for pairs of conceptually matched objects. Administered according to standard instructions. Together with Visual Matching this test is used to derive the WJ-III COG Processing Speed cluster score.

Cross Out. This test is taken from the Woodcock-Johnson Psycho-Educational Battery-Revised (Woodcock & Johnson, 1989) and requires testees to cross out 5
matching patterns in rows of distracting patterns. The standard 3 minute time limit was applied.

Tests of Gv

Block Design. Described in Study 1.

Spatial Relations. Described in Study 1.

Picture Recognition. This test from the WJ-III COG requires the recognition of a subset of previously presented pictures within a field of distracting pictures.

Additional abilities tests.

Raven’s Advanced Progressive Matrices-Short Form (APM-SF). A brief 12 item version of this test (Bors & Stokes, 1998) was used as a measure of fluid reasoning ability (Gf). A 15 minute maximum time limit was allowed.

Visual-Auditory Learning (VAL). From the WJ-III COG this test requires testees to learn, store and recall a series of visual-auditory associations. This test was included as an indicator for long-term storage and retrieval (Glr).

Customized Computer Tests

Symbol Digit. Described in Study 1.

Space Code Game. This test has the same layout and dimensions as Space Code (described in Study 1) with the addition of various computer-game-like elements. 1) Space ships to be destroyed were programmed to move according to a random movement algorithm; however, no aiming was required. 2) A timer was situated in the top right hand corner of the screen. The timer began at 120 seconds and proceeded downwards to indicate how much time was left. Additionally a verbal warning was provided at 90, 60, 30, 10 and 5 seconds. This was by way of male pitched synthesized speech provided under Windows XP
3) A scoring system was implemented and a cumulative total was presented in the top left of the screen. Bonus stars were also awarded at predetermined score levels and these were presented above and below the score. Each star indicated that a larger number of points would be awarded for each ship destroyed. The number of points awarded for each ship was displayed just to the left of the mouse response grid. All of these elements were described in an introduction screen prior to commencement of the game. 4) In addition to the laser and explosion sounds accompanying a correct response the score was incremented in steps and associated with a high pitched rising sound. 5) Incorrect responses were accompanied by an equivalent score decrement accompanied by a lower pitched descending tone. Additionally, to make incorrect responses more aversive, the response grid was also deactivated during the time it took for the score to decrease. This meant that the participant would have to wait for a short period before being able to fire again. It was hoped that this would discourage the rapid firing strategies used by some participants in Study 1. 6) Background music was played during the game to further enhance the game-like feel of the task. The dependent variable was the total number of correct responses.

**Guided Mouse-Grid Speed.** This test used the same response grid layout as used in Mouse-Grid Speed (described in Study 1). However, to improve on the task used in the first study, this test provided a more extensive demonstration of the task and patterns were additionally indicated by the response buttons lighting up. Participants were told that their task was to click on patterns as fast as they could for 15 seconds and that there would be an untimed practice run preceding each of four patterns in which they could learn the pattern. They were also instructed that the next number in each sequence would additionally be indicated by the corresponding button lighting up (this was intended to remove the
requirement of holding the pattern sequence in memory). The experimenter then performed a demonstration trial with the participant watching. The four numerical sequences completed by participants were \([1,2,3,4,5,6,7,8,9], [1,9,7,3,5], [2,6,8,4,5], \) and \([5,1,5,2,5,3,5,6,5,9,5,8,5,7,5,4] \). If the wrong button was clicked the current number button simply remained lit up. The dependent variable was the total number of correct clicks across all four patterns.

All computerized tests were run on a Pentium 4 class computer (2.4 MHz) with a Microsoft optical Intellimouse which was brand new at the start of the experiment. Display was via a 19 inch monitor with a vertical refresh rate of 100 Hz and a screen resolution of 1280 x 1024 pixels.

Procedure

Participants were tested individually in a quiet room and asked to concentrate as best as they could for each individual test. It was further emphasized that it was important for them to do their best on each test so that an accurate comparison could be made across their performances on all tasks. Test were administered to all participants in the following order: Guided Mouse-Grid Speed; Space Code Game; Digit Symbol; Visual Matching; Cross Out; Decision Speed; Symbol Digit; APM-SF; Visual-Auditory Learning; Block Design; Spatial Relations; Picture Recognition; Space Code Game. The second administration of Space Code Game was to provide an estimate of test-retest reliability. Sessions took approximately 2 hours to complete.

Study 2 - Results

Descriptive Statistics

Descriptive statistics for all tests are presented in Table 4.
Table 4

Descriptive statistics for ability test scores and computer based mouse response test scores (N=61).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>SD</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digit Symbol</td>
<td>92.8</td>
<td>15.0</td>
<td>59</td>
<td>123</td>
</tr>
<tr>
<td>Visual Matching&lt;sup&gt;a&lt;/sup&gt;</td>
<td>556.0</td>
<td>24.5</td>
<td>507</td>
<td>625</td>
</tr>
<tr>
<td>Decision Speed&lt;sup&gt;a&lt;/sup&gt;</td>
<td>538.8</td>
<td>18.8</td>
<td>496</td>
<td>586</td>
</tr>
<tr>
<td>Cross Out&lt;sup&gt;a&lt;/sup&gt;</td>
<td>524.2</td>
<td>9.0</td>
<td>500</td>
<td>542</td>
</tr>
<tr>
<td>Block Design</td>
<td>52.0</td>
<td>9.6</td>
<td>28</td>
<td>67</td>
</tr>
<tr>
<td>Spatial Relations&lt;sup&gt;a&lt;/sup&gt;</td>
<td>515.5</td>
<td>7.4</td>
<td>501</td>
<td>533</td>
</tr>
<tr>
<td>Picture Recognition&lt;sup&gt;a&lt;/sup&gt;</td>
<td>511.8</td>
<td>6.8</td>
<td>498</td>
<td>526</td>
</tr>
<tr>
<td>Visual Auditory Learning&lt;sup&gt;a&lt;/sup&gt;</td>
<td>504.1</td>
<td>8.8</td>
<td>484</td>
<td>527</td>
</tr>
<tr>
<td>APM-SF</td>
<td>7.0</td>
<td>2.9</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>Symbol Digit</td>
<td>89.4</td>
<td>15.6</td>
<td>62</td>
<td>132</td>
</tr>
<tr>
<td>Space Code Game</td>
<td>62.6</td>
<td>7.6</td>
<td>47</td>
<td>85</td>
</tr>
<tr>
<td>Guided Mouse-Grid Speed</td>
<td>177.1</td>
<td>25.3</td>
<td>122</td>
<td>256</td>
</tr>
</tbody>
</table>

<sup>a</sup>Scores provided are Rasch scaled W scores centred on 500 provided by the WJ-III CompuScore software package.

Correlations

The test-retest correlation for Space Code was .84, indicating acceptable reliability for this measure. There were no significant relationships between age and any of the test variables. However, gender was significantly related to Cross Out (<i>r</i> = .30, <i>p</i> = .017), Visual Matching (<i>r</i> = .27, <i>p</i> = .039), Guided Mouse-Grid Speed (<i>r</i> = .48, <i>p</i> < .001), Space Code Game (<i>r</i> = .31, <i>p</i> < .017) and APM (<i>r</i> = .31, <i>p</i> < .017). The positive relationships indicating that the male sample performed
somewhat better on these tests than females. These results suggest the smaller male sample may not be matched evenly for abilities with the female sample. Because the APM test is widely considered as an indicator of general ability, mean gender differences on this test were examined. Male scores ($M = 8.08$, $SD = 2.62$) were found to be significantly higher than female scores ($M = 6.30$, $SD = 2.86$), $t(59)= 2.46$, $p = .017$, $d = .64$, an outcome consistent with much work by Lynn (Irwing & Lynn, 2005). The suggestion that males in this sample were of higher ability is further supported by the finding that the usually reported female superiority for Digit Symbol performance (Nicholas R. Burns & Nettelbeck, 2005) was not observed in this sample ($r = .08$, $p = .534$).

Inter-correlations for Gs and Gv marker tests and CBMR tests are presented in Table 5. As expected the PP format Gs marker tests were all moderately to strongly correlated amongst themselves but not as strongly with the Gv tests. However, looking at the relationships amongst the three Gv marker tests it can be seen that although Block Design and Spatial Relations share considerable variance with each other, Picture Recognition is not significantly correlated with either of these tests. This suggests that Picture Recognition was not a good indicator of Gv within the present sample. Looking at the CBMR tests it can be seen that there is a positive manifold with all of the other ability tests and particularly with the Gs marker tests. This supports the interpretation of these tasks sharing common ability requirements with PP format Gs measures.
### Table 5

**Correlation matrix for ability test scores and computer based mouse response tests scores.**

<table>
<thead>
<tr>
<th>Variable</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Digit Symbol</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Visual Matching</td>
<td>.56''</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Decision Speed</td>
<td>.38''</td>
<td>.55''</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Cross Out</td>
<td>.45''</td>
<td>.45''</td>
<td>.53''</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Block Design</td>
<td>.34''</td>
<td>.29`</td>
<td>.32`</td>
<td>.36''</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Spatial Relations</td>
<td>.19</td>
<td>.35''</td>
<td>.07</td>
<td>.25`</td>
<td>.51''</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Picture Recognition</td>
<td>.14</td>
<td>-.02</td>
<td>.39''</td>
<td>.15</td>
<td>.14</td>
<td>-.12</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Symbol Digit</td>
<td>.67''</td>
<td>.35''</td>
<td>.38''</td>
<td>.57''</td>
<td>.28</td>
<td>.31`</td>
<td>.29`</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Space Code Game</td>
<td>.54''</td>
<td>.60''</td>
<td>.45``</td>
<td>.48''</td>
<td>.24</td>
<td>.37``</td>
<td>.12</td>
<td>.64''</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>10. Guided Mouse-Grid Speed</td>
<td>.37''</td>
<td>.45''</td>
<td>.35''</td>
<td>.48''</td>
<td>.14</td>
<td>.18</td>
<td>-.02</td>
<td>.44</td>
<td>.58''</td>
<td>-</td>
</tr>
</tbody>
</table>

* ` p < .05, ** p < .01
Exploratory Factor Analysis (EFA)

To further assess the construct validity of both the Symbol Digit and Space Code Game tests EFA procedures were carried out with these two tests and the Gs and Gv marker tests. In order to balance the number of measures for each constructs, Visual Matching and Decision Speed were used to derive the single Gs composite measure as indicated for the WJ-III COG (WJ-III Gs). Gs marker test were thus Digit Symbol, Cross Out and WJ-III Gs. The Gv construct was originally intended to be represented by Block Design, Spatial Relations and Picture Recognition. However, considering the very low correlations that Picture Recognition had with the other two tests, only the first two tests were included in the EFA. As in Study 1 PAF was used because the primary goal was to identify common factors and the number of factors to extract was decided by inspection of the scree plot (Cattell, 1966), parallel roots analysis (Montanelli & Humphreys, 1976), and the theoretical issues relevant to the present study.

Inspection of the scree plot indicated one or perhaps two factors. However parallel roots analysis indicated a one factor solution was appropriate. Extracting just the one factor resulted in a factor with high loadings (minimum .69) for all tests except Block Design and Spatial Relations which had loadings on this factor of .47 and .43, respectively. Considering this pattern of loadings and the theoretical aim of delineating a Gs and Gv factor it was decided that a two factor solution would be appropriate although the two factors were likely to be quite substantially correlated.

A two factor PAF solution was submitted to Promax rotation (Kappa = 4) and resulted in two correlated factors ($r = .52$). This solution accounted for 54.5% of the
Table 6

Structure Matrix for Principal Axis Factoring of ability test scores (Gs and Gv) and computer based mouse response test scores with Promax Rotation (Kappa = 4)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Processing Speed</th>
<th>Visual-Spatial</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(Gs)</td>
<td>(Gv)</td>
</tr>
<tr>
<td>Digit Symbol</td>
<td>.752</td>
<td>.370</td>
</tr>
<tr>
<td>WJ-III Processing Speed</td>
<td>.695</td>
<td>.430</td>
</tr>
<tr>
<td>Cross Out</td>
<td>.673</td>
<td>.430</td>
</tr>
<tr>
<td>Block Design</td>
<td>.404</td>
<td>.818</td>
</tr>
<tr>
<td>Spatial Relations</td>
<td>.368</td>
<td>.614</td>
</tr>
<tr>
<td>Symbol Digit</td>
<td>.798</td>
<td>.380</td>
</tr>
<tr>
<td>Space Code Game</td>
<td>.784</td>
<td>.392</td>
</tr>
</tbody>
</table>

Note. Bold numbers indicate loadings > .60

variance. The structure matrix of factor loadings is presented in Table 6. The Pattern Matrix indicated almost identical results. Using the marker tests as a primary guide, it can be seen that the first factor is strongly loaded by the three traditional Gs markers and the second factor is strongly loaded by the two Gv tests. This replicates the theoretical measurement framework found in Study 1, although the two factors are more highly correlated within this sample. Looking at the loadings for the CBMR tests, both Symbol Digit and Space Code Game have loadings patterns of similar magnitude to the paper-and-pencil Gs marker tests. This supports the interpretation of both Symbol Digit and Space Code Game as Gs measures. Contrary to results for the prototype Space Code test in Study 1, Space Code Game does not show balanced loadings across both Gs and Gv factors.
Correlations with Composite Measures and Other Ability Tests

To further assess the construct validity (concurrent and discriminant) of the CBMR tests, correlations with composite Gs and Gv scores (as described above) and the two other ability tests, Raven's APM-SF and VAL, were calculated. To enable a direct comparison with a PP speeded coding task, the same correlations were calculated for the Digit Symbol test. To avoid the problem of part-whole correlations the Gs composite was calculated using all of the paper-and-pencil Gs markers except for Digit Symbol. These correlations are shown in Table 7. It can be seen that the CBMR tests show a very similar pattern of correlations with the composites and ability tests as Digit Symbol. Guided Mouse-Grid Speed shows the simplest relationship with the Gs composite in that it is not significantly related to Gv or VAL. It also appears that Symbol Digit and Space Code Game may to some extent tap an associative memory ability more than does Digit Symbol, as evidenced by these tests having similar but small relationships with the VAL test.

Although the CBMR tests do not share a PP methodology with the tests comprising the Gs composite both of these test are in fact slightly more strongly correlated with these than Digit Symbol is. All of the tests share a similar relationship with the APM.
Table 7

Correlations between computer based mouse response test scores and composite Gs and Gv, Raven’s Advanced Progressive Matrices-Short Form (APM-SF) and Visual-Auditory Learning (VAL) test scores.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Gs</th>
<th>Gv</th>
<th>APM-SF</th>
<th>VAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digit Symbol</td>
<td>.57**</td>
<td>.31*</td>
<td>.38**</td>
<td>.09</td>
</tr>
<tr>
<td>Guided Mouse-Grid Speed</td>
<td>.53**</td>
<td>.18</td>
<td>.46**</td>
<td>.07</td>
</tr>
<tr>
<td>Symbol Digit</td>
<td>.63**</td>
<td>.34**</td>
<td>.45**</td>
<td>.25*</td>
</tr>
<tr>
<td>Space Code Game</td>
<td>.67**</td>
<td>35**</td>
<td>.49**</td>
<td>.24</td>
</tr>
</tbody>
</table>

*p < .05, **p < .01

Discussion

Results across the two studies considered together suggest that a computer-game-like test of Processing Speed (Gs) may provide a viable alternative to traditional PP measures. In Study 1 it was found that the simple computerized coding test loaded uniquely on the Gs factor. These results support the previous finding by McPherson and Burns (2005) who also found that this test loaded on a Gs factor independently of a second factor loaded by reaction time tasks. In Study 1 it was also found that Space Code, as well as loading on the Gs factor, additionally loaded on the Gv factor. This suggested the possibility that the use of more complex visual stimuli introduces performance variance associated with visual abilities. However, in Study 2 this result was not replicated and instead Space Code Game had a loading on the Gv factor similar to the Gs marker tests. It seems unlikely to us that this resulted from the addition of the more dynamic game-like elements to the task as these did not have any effect on the nature of the visual stimuli. We emphasize that further research is necessary to more fully
assess the degree to which Gs tests using game-like stimuli may tap both and Gs
and Gv related processes.

Despite this issue, the potential of a game-like test of Gs was further evident in
the results from Study 2. Similar correlations between Gs marker tests and Space
Code to those obtained in Study 1 were found but using a larger range of marker
tests. Both Space Code Game and the simpler coding task showed discriminant
validity in that these were not strongly related to a measure of associative learning
or delayed recall (Glr), although the relationships were slightly higher than those
obtained between Digit Symbol and this measure. This suggests that the tests
primarily measure the speed of coding and not the ability to memorise
associations between the numbers and spaceships or symbols. Space Code also
shows signs of being psychometrically reliable with a test-retest correlation of .84.

Interestingly, all of the computer based tests had a moderate to strong
relationship with Raven’s APM. It should also be pointed out that Guided Mouse-
Grid Speed showed a substantial relationship with the Gs composite measure and
Raven’s APM. Thus, this task, which was primarily designed to provide a measure
of how quickly people could utilise the mouse response method, may in fact tap
cognitive processes more central to traditional abilities tests. The finding that
Space Code showed a substantial relationship with the Raven’s APM, a test
considered by many researchers to be central within the spectrum of human
cognitive abilities (Carpenter, Just, & Shell, 1990; Prokosch, Yeo, & Miller, 2005)
further supports the notion that game-like tests can be developed to supplement
traditional cognitive abilities tests.

Although we are optimistic about the potential for developing computer-game-
like tests of cognitive abilities, a number of important issues remain to be
investigated. These include the nature of any possible gender bias effects
associated with various game elements. A recent survey of computer-game users suggests that there are differences in the preferences of males and females in regards to the types of games they prefer and which elements are most important to them within gaming environments (Wood et al., 2004). In Study 1, gender and Digit Symbol performance were negatively correlated, replicating the commonly found phenomenon of faster female performance on this test, but gender and Space Code were positively correlated. Further research is necessary to ascertain whether this result is stable and if so what might underlie differential performance on these tasks. Other important issues include whether noisy game-like environments (aurally and visually) affect particular individuals or groups differently and whether different cover stories are more engaging for some people. Nevertheless, we believe that these issues are not seriously detrimental to the development of computer-game-like assessment tools and that there are many potential uses for such tools.

As computer technologies continue to progress, it would seem a wasted opportunity for psychological assessment to ignore new possibilities that are created by this ever expanding set of tools. Indeed, it would seem that many researchers are already beginning to venture out of traditional territory with the development of complex simulation assessments (Kröner et al., 2005). We believe that computer-game-like assessments could similarly provide a useful and interesting addition to the types of assessments traditionally used in abilities research and practice. This process should involve assessing whether computer-game-like tests can be designed to measure already identified constructs, as we have attempted to do in the present paper, as well as whether there are in fact new constructs tapped by game-like environments.
Assessing the validity of computer-game-like tests of processing speed and working memory

Jason McPherson and Nicholas R. Burns (accepted, March 2008)

Behavior Research Methods

Statement of Contributions

Jason McPherson (Candidate)
I was responsible for the conception and primary authorship of the paper. I was responsible for the development and programming of software based assessment tools and collection of all data in Study 1. Data in Study 2 was collected by Ms Ertimis Eshkevari under joint supervision of myself and the co-author. I conducted the statistical analyses independently with advice from the co-author. I was corresponding author and primarily responsible for responses to reviewers and revisions to the paper.

Signed…………………………………….Date…………………………………….

Nicholas R. Burns (Co-author)
I was the supervisor (advisor) for the research programme that lead to this publication. In terms of conceptualisation of the programme, there was extensive and ongoing collaboration between Mr. McPherson and me in developing the direction of the research. The realization of the program, specifically, development of the computer-game-like and other computerized tests, data collection, and
analyses, were the work of Mr. McPherson. I had an advisory role with respect to
selection of the test battery used and on the direction and specifics of the data
analyses.
Mr. McPherson was responsible for writing this paper; my role was to comment on
drafts, make suggestions on the presentation of material in the paper, and to
provide editorial input. I also provided advice on responding to comments by the
journal reviewers and editor.
I hereby give my permission for this paper to be incorporated in Mr. McPherson’s
submission for the degree of PhD in the University of Adelaide.

Signed…………………………………….Date…………………………………….
Abstract

Processing speed (Gs) and working memory (WM) tasks have received considerable interest as correlates of more complex cognitive performance measures. However, Gs and WM tasks are often repetitive and rigidly presented and may therefore be confounded with motivation and anxiety effects. In an effort to address this, we assessed the concurrent and predictive validity of computer-game-like tests of Gs (Space Code) and WM (Space Matrix) across two studies. In Study 1, within a university sample (N = 70), Space Matrix exhibited concurrent validity as a WM measure while Space Code appeared to be a mixed ability measure. In Study 2, within a school aged sample (N=94), Space Matrix exhibited concurrent validity as well as predictive validity (as a predictor of school grades), while results for Space Code were less encouraging. Relationships between computer-game-like tests and gender, handedness, and computer game experience are also discussed.
Introduction

One of the key challenges within cognitive abilities research has been to identify simple cognitive tasks that can explain individual variation in more complex cognitive performance criteria such as composite ability ('IQ') measures, abstract reasoning (fluid intelligence; Gf\(^1\)), and academic achievement (I.J. Deary, 2001; Engle et al., 1999; Jensen, 1998; McGrew, 2005; Schweizer, 2005). Most of the simple cognitive tasks identified within this research tradition were originally developed as experimental measures of underlying ‘information processing’ components derived from various models of human cognition (see Floyd, 2005, for a review). Two information processing components that have received considerable interest and support as correlates of more complex cognitive performance measures are processing speed and working memory (Conway et al., 2002; Engle et al., 1999; Fry & Hale, 1996; Luo et al., 2006; McGrew, 2005).

Processing speed (Gs) is measured with tasks requiring rapid performance of simple cognitive operations. Items comprising Gs tasks are so uniformly simple that people generally only differ in their speed of responding rather than their accuracy (Carroll, 1993; Horn & Noll, 1997). In contrast, working memory (WM) measures are primarily measures of accuracy, requiring the maintenance of “memory representations in the face of concurrent processing, distraction, and/or attention shifts” (Conway et al., 2002, p. 164). What Gs and WM tasks share are their relatively homogeneous measurement units: Gs being measured by the number of uniformly simple items completed within a given time limit; WM being measured by the number of uniformly simple representations maintained.

---

\(^1\) Ability nomenclature and abbreviations are those used within Cattell-Horn-Carroll (CHC) Theory (McGrew, 2005; McGrew & Flanagan, 1998)
The precise relationship between measures of Gs and more complex ability measures has been investigated and debated for decades (Burns & Nettelbeck, 2003; Danthiir et al., 2005; Jensen, 1982, 1998; Nyborg, 2003; Stankov & Roberts, 1997). Similarly, a voluminous amount of research and commentary has focused on the relationship between WM and more complex ability measures (Ackerman et al., 2005; Gathercole & Pickering, 2000; Just & Carpenter, 1992; Kane, Hambrick, & Conway, 2005; Kyllonen & Christal, 1990; Oberauer, Schulze, Wilhelm, & Süß, 2005; Wickelgren, 1997). More recently researchers have begun to investigate the complex inter-relationships between Gs, WM, Gf, and broader composite ability measures (Conway et al., 2002; Fry & Hale, 1996, 2000; McGrew, 2005; Süß, Oberauer, Wittmann, Wilhelm, & Schulze, 2002). The relatively consistent findings of substantial relationships amongst these variables has lead some researchers to propose Gs and WM tasks as viable assessment tools in educational and clinical settings (Lengenfelder et al., 2006; Luo et al., 2006).

Luo, Thompson and Detterman (2006) argue that Gs and WM tasks require little formal instruction and do not seem to require the higher order processes thought to underlie complex reasoning. Additionally, the relatively homogeneous nature of items in Gs and WM tasks means that item banks are easily generated and are also amenable to experimental manipulation. This ease of manipulation also extends to content and modality features of the stimuli used such that Gs and WM tasks can be constructed using verbal, numerical or spatial stimuli and presented visually or aurally in many cases. Although there is some evidence that domain specific processes contribute to Gs and WM task performance, research also suggests that the same more general underlying processes are measured at a broader level (McGrew, 2005; Miyake, 2001). However, despite the aforementioned advantages and the finding that Gs and WM measures were
predictors of scholastic achievement, Luo, Thompson and Detterman also acknowledge that these tasks tend to be rigidly presented and often appear ‘impersonal’. Consequently, they suggest that Gs and WM measures may be more confounded with psychological effects such as anxiety and weakened motivation (p. 81).

We, and numerous other researchers, have previously suggested that computer-game-like tasks may provide a less anxiety provoking and more motivating environment for those being tested (McPherson & Burns, 2005, 2007; Porter, 1995; Washburn, 2003). This may especially be the case for children who are increasingly familiar with computers via their experience with computer games (Gentile & Walsh, 2002; Yelland & Lloyd, 2001). Computer games incorporate a range of unique structural features that act to engage and motivate players, such as real time positive and negative feedback (via graphics, sounds, and scoring), realistic graphics and sounds, and player advancement through levels (Wood et al., 2004). By incorporating such features, computer-game-like Gs and WM tasks could potentially maintain many of the advantages inherent in such measures, while also addressing the concerns raised by Luo, Thompson and Detterman (2006).

Alternatively, introducing game-like elements might distract those being tested to the point that the construct of interest is no longer reliably measured. Indeed, as Gs and WM measures are also thought to involve strong attentional demands (Conway et al., 2002; Horn & Blankson, 2005; Schweizer & Moosbrugger, 2004)- and to the extent that such tests may also measure the ability to concentrate on somewhat monotonous tasks- altering motivationally sensitive aspects of the tasks could impact on their psychometric properties. However, we believe the development of game-like tests may help to investigate the separate contributions
of speed and attention in Gs tasks once their psychometric properties have been established. In our previous research we have attempted to address this problem, and the additional ergonomic issue of using a computer mouse response method, by developing a computer-game-like test of Gs in distinct stages.

In a preliminary study we developed a simple speeded coding task, using a mouse response method and non game-like stimuli, and found that the task loaded on a Gs factor (McPherson & Burns, 2005). In two subsequent studies we investigated the impact of introducing game-like features (McPherson & Burns, 2007). In the first of these studies we introduced complex game-like graphics as stimuli and found that the task shared variance with both Gs and Visuospatial (Gv) tests. In the final study we added further dynamic game-like features (e.g., scoring, feedback and additional graphics and sound features) and assessed the task within a larger range of ability tests. The task (Space Code) was found to load strongly on a Gs factor and only minimally on a Gv factor. We concluded that Space Code may prove to be a viable measure of Gs, albeit requiring further research, but there was also some evidence suggesting the task may have a more complex psychometric profile.

The development of Space Code was partially suggested by the observation that many computer games inherently require rapid responding of a similar nature to many Gs tasks. Development of Space Code was also aided by the fact that Gs tasks can be readily adapted to different stimuli classes, including game-like graphics in this instance. In a similar manner it could also be argued that many modern games require the maintenance of memory representations in the face of other distracting tasks – much like WM tasks do – such as recalling where objects, doorways, or other rewards are located. Additionally, as with Space Code, there
appears to be no reason why game-like graphics cannot be used as stimuli within a WM task.

The present paper describes two related studies. The overall aim of both studies was to further assess the validity of Space Code, as a measure of Gs, as well as a new computer-game-like WM task, Space Matrix. Space Matrix was developed by using the core task of Space Code and adding a simultaneous task based on the Dot Matrix task developed by Miyake, Friedman, Rettinger, Shah, & Hegarty (2001). In Study 1 the concurrent validity of both tests was assessed, in relation to established Gs, WM and Gf measures, within a university student sample. In Study 2, concurrent validity was assessed as in Study 1, and predictive validity was assessed, using school subject grades, within a sample of school aged children. Additionally, each study also assessed relationships with gender, handedness, and computer game experience measures.

Study 1 - Method

Participants

There were $N = 70$ participants. All participants were Level I Psychology students from the University of Adelaide who received course credit for their participation. The mean age for the whole sample was 19.6 years ($SD = 4.00$), 40 were female (Mean age = 19.20 years, $SD = 2.37$) and 30 were male (Mean age = 20.13 years, $SD = 5.48$).

Materials

Non-Ability Measures

Participants were asked to indicate their age (in years), their gender, and whether they considered themselves predominantly right handed, somewhat
ambidextrous, or predominantly left handed (scored 1, 2 and 3, respectively). Additionally, there were three questions aimed at measuring computer game experience. The questions were: 1) On average how many hours do you play computer games each week; 2) On average how many hours do you play computer games using a mouse each week; and 3) At any stage in the past what was the maximum number of hours you played computer games in a week. Participants were able to freely estimate the number of hours for each question.

Tests of Processing Speed (Gs)

Digit Symbol. This test from the Wechsler Adult Intelligence Scale (WAIS-III) requires the filling in of blank cells according to a key provided at the top of each test sheet. The standard two minute time limit was used.

Visual Matching. This test, taken from the Woodcock-Johnson III Tests of Cognitive Abilities (WJ-III COG; Woodcock, McGrew, & Mather, 2001) requires searching for and circling matching pairs of numbers. The standard three minute time limit was used.

Decision Speed. This test from the WJ-III COG requires searching for pairs of conceptually matched objects. The standard three minute time limit was used.

Tests of Working Memory/ Fluid Intelligence (WM/Gf)

Picture Swaps. This test was adapted from the swaps test (Stankov, 2000) which has been used as a marker test for working memory (WM) in previous research (Schweizer & Moosbrugger, 2004). This test was a computerised version in which alphabetical symbols were replaced with colourful cartoon pictures of animals and everyday objects. Pictorial stimuli were used so as to make the test more visual in nature. Each item consisted of a ‘swap’ instruction screen and a response screen. Each instruction screen consisted of three pictures presented in
a row with written instructions below asking participants to swap the order of pictures mentally. Participants could proceed to the response screen by clicking a button whenever they were ready. The response screen presented all six possible re-orderings of the pictures and participants chose the order they thought was correct by selecting it and left clicking the mouse. Participants had to correctly complete two practice questions with one swap and two with two swaps prior to commencing the actual test. There were 16 items in total and three levels of difficulty; items 1-2 required two swaps, items 3-8 required three swaps, and items 9-16 required four swaps. An incremental number of items at each successive swap level was used as it was expected that a university sample would be better differentiated by more difficult items. There were no time limits imposed on any section of the test. The dependent variable was the total number of correct responses. Cronbach’s alpha was .78.

**Dot Matrix.** This test was based on the Dot Matrix task (Miyake et al., 2001) which was originally based on the Spatial Verification Task (Law, Morrin, & Pellegrino, 1995). This task has been used as a test of Visuospatial WM in a number of studies investigating the structure of WM (Colom, Abad, Rebollo, & Chun Shih, 2005; Colom & Shih, 2004; Law et al., 1995). In this computerised version of the task participants viewed a matrix equation (see Figure 1) and then verified whether the equation was true or false by left mouse clicking a TRUE or FALSE button. If an incorrect response was made a pop-up warning (“No, look again closely”) was displayed and the equation was repeated. If no response was made within eight second was presented. This was repeated every eight seconds if a response was not made. Following a correct response a 5 x 5 grid with a dot located in one of the grid squares was presented for 1500 ms. After a series of equation and grid displays (ranging from 2 to 5) participants were asked to indicate
locations within the 5 x 5 grid. Participants responded by clicking on an empty 5 x 5 grid to place dots where they recalled their locations and then left clicking the ‘Click to enter answer’ button when they were happy with their responses. The computer program enabled participants to change their responses by clicking on a square to deselect it and they were able select any number of grid positions up to and including the number of grid positions presented. If more than the number of grid positions were selected a pop-warning (“Too many selected”) persisted until a suitable number of grid positions were deselected. Prior to commencing the actual test participants had to successfully complete two practice questions consisting of two equation-grid pairs. The test had four levels (with 2, 3, 4 and 5 equation-grid pairs) with three items at each level. Each successive level required a larger number of grid locations be held in memory. For example, a level 5 item would present 5 equations each followed by a different grid location to be remembered.
There were no time limits imposed on any section of the test beyond those already mentioned. The dependent variable was the number of correctly recalled grid locations. Cronbach’s alpha was .79.

Raven’s Advanced Progressive Matrices Short Form (APM-SF). This was a computerised 12 item version of the original test (Raven, Raven, & Court, 1998). The 12 items used were the same subset validated by Bors and Stokes (1998) for use as a brief form. Items were presented by displaying the original stimuli on a computer screen with the response options numbered from 1 to 8. Responses were made by using the computer mouse to select the number corresponding to the option they thought was correct and then clicking an ‘Enter’ button to enter their response. The ‘Enter’ button was included to allow participants to review their answer first. This was included as a measure of fluid reasoning ability (Gf). A 15 minute timer on screen was provided as a guideline but respondents were told they could continue if they desired. Cronbach’s alpha was .70.

Computer-Game-Like Tests

Space Code. This was the same test as ‘Space Code Game’ described in a previous validation study (McPherson & Burns, 2007). The screen layout for this test consisted of a spaceship cockpit with a numerical response grid situated in the centre of a control panel below a window view looking out into space ahead (Figure 2). The grid was a 3 x 3 numerical grid (6.8 cm x 6.2 cm) numbered from 1 to 9 beginning in the top left grid position. The grid sections were dark blue with white digits and each number’s background turned light blue when the cursor was placed over it. To make the grid easier to use the cursor was restricted such that it could not be moved outside of the grid area. Each item consisted of a single spaceship (approximately 5cm x 5cm) appearing in the window view of the cockpit.
To destroy spaceships appearing in the centre of the cockpit window view the correct number laser must be fired on the numeric grid. The correct laser number is determined by matching the spaceship according to the code table at the bottom of the screen. The correct laser for this item is 7.

At the bottom of the cockpit control panel nine spaceships were each presented with a single digit placed directly above. Participants were instructed by on-screen text that correct responses (i.e., the destruction of each target spaceship) required the clicking/firing of the corresponding number as indicated above each spaceship at the bottom of the screen. The next item was not presented until the correct laser was fired. Seven practice items were provided before the test began. Participants were instructed to be as fast and accurate as they could in an effort to destroy as many spaceships as possible in two minutes.

Space Code also included various computer-game-like elements which were described in an introduction screen prior to commencement of the game. The game-like elements were as follows. Spaceships were programmed to move according to a random movement algorithm and zoomed in from a distant...
perspective point; however, no aiming was required. When the correct grid number was clicked, two laser bullets accompanied by a ‘laser’ sound, were seen shooting towards the target spaceship leading to a visible and audible explosion. When an incorrect grid number was clicked a low pitched descending tone was heard, accompanied by voice synthesised speech (“wrong laser”; Windows XP, Microsoft Sam) and the response grid was deactivated for approximately 500ms. This was intended as an aversive consequence for rapid inaccurate firing.

A scoring system was implemented beginning at 150 points for each correctly destroyed ship and a cumulative total was presented in the top left of the screen. Up to ten bonus stars were also awarded at predetermined score levels (1000, 5000, and every multiple of 10000 up to 80000) with these presented above or below the score and accompanied by a ‘whooshing’ sound when awarded. Each star also indicated that a larger number of points would be awarded for each ship destroyed; there was a 150 point increase for each bonus star. The number of points awarded for each ship was displayed just to the left of the mouse response grid. In addition to the lasers and explosion sounds accompanying a correct response the score was incremented in visible steps and associated with a high pitched ascending tone. Incorrect responses were accompanied by an equivalent visible score decrement. This was further intended to make incorrect responses more aversive. A timer was situated in the top right hand corner of the screen. The timer began at 120 seconds and proceeded downwards to indicate how much time was left with each second decrement accompanied by a faint beep sound. Additionally, a verbal warning (Windows XP, Microsoft Sam) was provided at 90, 60, 30, 10 and 5 seconds. Background music was played during the game to further enhance the game-like feel of the test. The dependent variable was the total number of correct responses made in two minutes. The number of spaceships
correctly destroyed in the first and second sixty second blocks of the test were recorded and correlated to provide a measure of split-half reliability ($r = .75, p < .001$).

**Space Matrix.** In this test participants were asked to destroy spaceships exactly as in Space Code while also monitoring where dots were located on the same type of 5 x 5 grid as used in Dot Matrix. These dot locations were described as indicating which ‘sector’ of space they were operating in and participants were instructed that they would have to report back to headquarters which sectors they had been operating in from time-to-time. The screen layout was the same as for Space Code with the addition of the sector grid appearing at intervals to the right of the numerical response grid on the cockpit control panel. Sectors were presented at seven second intervals, accompanied by a beep at onset and displayed for 2.5 seconds only. After a sequence of sectors had been displayed (ranging from 2 to 4) participants were asked to indicate which sectors they recalled operating in. Participants responded by placing dots on an empty grid, as in Dot Matrix, and then clicking an ‘Enter Sectors’ button when they were happy with their responses. Sectors could be deselected by clicking on them again. If a smaller number of sectors were selected than displayed a pop-up warning (“Too few selected, guess if you can’t remember”) was presented. If more than the number of displayed sectors was selected another pop-up warning (“Too many selected”) was presented. Difficulty levels of the task were manipulated by increasing the number of sectors displayed before asking for sector reports. Participants completed two practice ‘runs’ with two sectors displayed and one practice run with three sectors displayed. Feedback was provided indicating how many sectors were correctly identified.
The test consisted of 36 sector locations that were presented in sequences of two, three, or four sectors; there were six sequences of two sectors, four sequences of three sectors, and three sequences of four sectors. A scoring system was implemented and presented as in Space Code but with a fixed 300 points scored for each correctly destroyed ship and 500 ‘Bonus’ points for each correctly reported sector. The bonus points were awarded after sectors were entered and displayed graphically with a scrolling score display and an accompanying sound rising in pitch. Some elements from the original Space Code were removed to decrease the number of extraneous stimuli as piloting suggested the dual task nature of the task already provided a significant amount of stimulation for players; the bonus star system was removed and background music was replaced by a constant low humming sound akin to space ship engine noise. Instructions emphasised that both destroying ships and reporting sectors accurately were important for the success of the mission. The primary dependent measure was the total number of correctly reported sectors out of the 36 presented throughout the test. The number of ships correctly destroyed was also obtained as a potential speed measure. The actual test took approximately 4.5 minutes to complete. Each sequence of sectors reported served as an ‘item’ and separate scores were recorded for each sequence to provide a measure of internal consistency. Cronbach’s alpha for these item units was .76.

Procedure

Participants were tested individually, or in pairs, in a quiet room and asked to concentrate as best as they could for each individual test. It was further emphasized that it was important for them to do their best on each test so that an accurate comparison could be made across their performances on all tasks. Test were administered to all participants in the following order: Non-Ability Measures;
Space Code; Space Matrix; Visual Matching; Decision Speed; Digit Symbol; Dot Matrix; Swaps; APM-SF Space Code; Space Matrix. A consistent test order was used in an effort to make any fatigue or motivational effects equal across all participants and to provide the largest possible gap between the first and second administrations of Space Code and Space Matrix, which were included to provide some estimate of task stability. Sessions took approximately 2 hours to complete.

All computerized tests were run on a Pentium 4 class computer (2.4 MHz) with a Microsoft optical Intellimouse. Display was via a 19 inch CRT monitor with a vertical refresh rate of 100 Hz and a screen resolution of 1280 x 1024 pixels.

Study 1 - Results

Descriptive Statistics

Descriptive statistics for all tests are presented in Table 1.1. Looking at the means for the two administrations of Space Code a substantial practice effect was observed, $t(65) = 23.74, p < .001, d = 2.29$ The effect size for the mean difference was approximately twice the size as was the case in a previous study (McPherson & Burns, 2007) which was probably due to the increased exposure to the codes gained in Space Matrix. The correlation between the two sets of test scores ($r = .75, p < .001$) suggested Space Code was reliable across the two administrations. There was a smaller but significant practice effect for Space Matrix, $t(65) = 5.26, p < .001, d = .52$ and the correlation between the two administrations was also quite strong for Space Matrix ($r = .77, p < .001$). The difficulty level for Space Matrix appeared to be quite well calibrated for this sample with the mean score being slightly larger than half of the maximum possible score of 36 for both administrations. There was also a significant practice effect for Space Matrix Speed, the speed measure taken from Space Matrix, $t(66) = 15.10, p < .001, d =
Table 1.1

Descriptive statistics for ability and computer-game-like tests.

<table>
<thead>
<tr>
<th>Test</th>
<th>Mean</th>
<th>SD</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digit Symbol</td>
<td>87.4</td>
<td>13.3</td>
<td>50</td>
<td>115</td>
</tr>
<tr>
<td>Visual Matching(^a)</td>
<td>559.3</td>
<td>26.0</td>
<td>510</td>
<td>625</td>
</tr>
<tr>
<td>Decision Speed(^a)</td>
<td>541.2</td>
<td>19.5</td>
<td>499</td>
<td>586</td>
</tr>
<tr>
<td>Picture Swaps</td>
<td>11.9</td>
<td>3.1</td>
<td>6</td>
<td>16</td>
</tr>
<tr>
<td>Dot Matrix</td>
<td>32.2</td>
<td>5.7</td>
<td>20</td>
<td>42</td>
</tr>
<tr>
<td>APM-SF</td>
<td>6.4</td>
<td>2.7</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>Space Code</td>
<td>61.0</td>
<td>9.1</td>
<td>40</td>
<td>90</td>
</tr>
<tr>
<td>Space Code(^b)</td>
<td>87.5</td>
<td>13.6</td>
<td>58</td>
<td>111</td>
</tr>
<tr>
<td>Space Matrix</td>
<td>19.3</td>
<td>6.2</td>
<td>8</td>
<td>34</td>
</tr>
<tr>
<td>Space Matrix(^b)</td>
<td>22.8</td>
<td>7.2</td>
<td>9</td>
<td>36</td>
</tr>
<tr>
<td>Space Matrix Speed</td>
<td>111.8</td>
<td>31.0</td>
<td>57</td>
<td>187</td>
</tr>
<tr>
<td>Space Matrix Speed(^b)</td>
<td>134.9</td>
<td>26.4</td>
<td>80</td>
<td>182</td>
</tr>
</tbody>
</table>

\(^a\) Based on Rasch scaled W scores centred on 500.

\(^b\) Scores from second administration of test.

Note. APM-SF is Raven’s Advance Progressive Matrices Short Form. Space Matrix Speed is the number of ships destroyed in Space Matrix.

.80, and there was a very strong relationship between the two administrations (\(r = .93, p < .001\)).

Test Intercorrelations

Correlations between all ability tests are presented in Table 1.2. Relationships between the Gs and WM/Gf marker tests were generally as theoretically expected. The three Gs marker tests, Digit Symbol, Visual Matching, and Decision Speed were all significantly correlated with each other as were the three WM/Gf tests, Picture Swaps, Dot Matrix, and APM-SF. Correlations between the two types of
marker tests were generally smaller but most were statistically significant. Both
computer-game-like tests had substantial correlations with a number of traditional
tests and these were of a similar magnitude to the correlations between the
marker tests themselves. The speed measure from Space Matrix (Space Matrix
Speed) was strongly related to Space Code but not with the primary Space Matrix
memory measure thus indicating independence of the speed and working memory
components of the test.

**Exploratory Factor Analyses (EFA)**

The tests included in EFA procedures were the three Gs markers (Digit Symbol,
Visual Matching and Decision Speed), the three WM/Gf measures (Picture Swaps,
Dot Matrix and APM-SF), and the two computer-game-like tests, Space Code and
Space Matrix. The APM-SF was expected to load with the WM tasks in the
absence of other Gf marker tests. Kaiser-Meyer-Olkin (.841) and Bartlett \(\chi^2 =
216.85, df = 28, p < .001\) statistics indicated the correlation matrix was suitable for
EFA. An initial Principal Components Analysis resulted in two components with
eigenvalues larger than unity (4.02 and 1.32 respectively) and a third eigenvalue
of 0.64. The first two components accounted for 50.2% and 16.5% of the total
variance respectively. However, to assess latent construct validity more
specifically subsequent analyses were conducted using Principal Axis Factoring
(PAF).

The number of factors to extract was assessed using multiple decision criteria
(Fabrigar et al., 1999; Henson & Roberts, 2006). These were the Scree test
(Cattell, 1966), Parallel Roots Analysis (Montanelli & Humphreys, 1976) and
Velicer’s Minimum Average Partial (Velicer, 1976) as well as consideration of
residual correlations and the theoretical issues relevant to the present study. PRA
was conducted for 1000 random permutations of the raw data using SPSS syntax
Table 1.2.

Correlation matrix for ability and computer-game-like tests.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Digit Symbol</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 Visual Matching</td>
<td>.57**</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 Decision Speed</td>
<td>.53**</td>
<td>.59**</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 Picture Swaps</td>
<td>.36**</td>
<td>.34**</td>
<td>.23</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 Dot Matrix</td>
<td>.39**</td>
<td>.46**</td>
<td>.43**</td>
<td>.54**</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 APM-SF</td>
<td>.30*</td>
<td>.24*</td>
<td>.11</td>
<td>.63**</td>
<td>.48**</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 Space Code(^a)</td>
<td>.45**</td>
<td>.48**</td>
<td>.37**</td>
<td>.37**</td>
<td>.41**</td>
<td>.42**</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>8 Space Matrix(^a)</td>
<td>.33**</td>
<td>.42**</td>
<td>.26</td>
<td>.54**</td>
<td>.66**</td>
<td>.53**</td>
<td>.51**</td>
<td>-</td>
</tr>
<tr>
<td>9 Space Matrix Speed(^a)</td>
<td>.29*</td>
<td>.27</td>
<td>.34**</td>
<td>.21</td>
<td>.15</td>
<td>.25</td>
<td>.79*</td>
<td>.15</td>
</tr>
</tbody>
</table>

\(^{*}\ p < .05, \ ** \ p < .01.

\(^a\) Scores are for first administration of test.

Note. APM-SF is Raven’s Advance Progressive Matrices Short Form.
by O’Connor (2000) and the MAP test was performed using SPSS syntax for extension analyses also by O’Connor (2001).

Visual inspection of the scree plot and both PRA and MAP criteria indicated two factors. A two factor PAF solution was obtained explaining 56.16% of the common variance and the two factors were then rotated using the oblique Direct Oblimin procedure (Delta = 0). The correlation between the two rotated factors was .50. There were no residual correlations larger than .10; the largest residual correlation was between Dot Matrix and Space Matrix ($r = .09$). Inspection of the Pattern Matrix (Table 1.3) indicated two quite distinct factors. Factor 1 was strongly loaded by all of the WM/Gf marker tests and Space Matrix. Factor 2 was strongly loaded by all of the paper-and-pencil Gs tests. Space Code was not strongly loaded on either factor. However, as the two factors were quite strongly correlated it is important to consider the Structure Matrix also (Henson & Roberts, 2006). It can be seen in Table 1.3 that Space Code was similarly correlated with both factors. Digit Symbol and Visual Matching were very strongly correlated with Factor 2 but also had substantial correlations with Factor 1. Similarly, Dot Matrix and Space Matrix were strongly correlated with Factor 1 but also had substantial correlations with Factor 2. Despite a degree of inter-relatedness between the factors being apparent, it seems reasonable to interpret the factors as representing the two hypothesised constructs; Factor 1 is thus interpreted as WM/Gf and Factor 2 as Gs. However, particularly considering Space Code’s shared loadings, it was decided that a Schmid-Leiman transformation (Schmid & Leiman, 1957) would further aid interpretation of these results.

The Schmid-Leiman transformation procedure enables the independent contributions of first and second order factors on the variables to be assessed
Table 1.3

Summary of two factor PAF solution.

<table>
<thead>
<tr>
<th></th>
<th>Pattern Matrix</th>
<th>Structure Matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Factor 1</td>
<td>Factor 2</td>
</tr>
<tr>
<td>Digit Symbol</td>
<td>.121</td>
<td>.636</td>
</tr>
<tr>
<td>Visual Matching</td>
<td>.068</td>
<td>.769</td>
</tr>
<tr>
<td>Decision Speed</td>
<td>-.121</td>
<td>.828</td>
</tr>
<tr>
<td>Picture Swaps</td>
<td>.755</td>
<td>.002</td>
</tr>
<tr>
<td>Dot Matrix</td>
<td>.580</td>
<td>.268</td>
</tr>
<tr>
<td>APM-SF</td>
<td>.837</td>
<td>-.148</td>
</tr>
<tr>
<td>Space Code</td>
<td>.375</td>
<td>.368</td>
</tr>
<tr>
<td>Space Matrix</td>
<td>.716</td>
<td>.108</td>
</tr>
<tr>
<td><strong>Eigenvalue</strong> a</td>
<td>3.04</td>
<td>2.76</td>
</tr>
<tr>
<td><strong>% of Variance</strong> b</td>
<td>38.01</td>
<td>34.52</td>
</tr>
<tr>
<td>(Cumulative %)</td>
<td>(56.10)</td>
<td>(56.10)</td>
</tr>
</tbody>
</table>

a Eigenvalues presented are for extracted factors after rotation.

b Indicates percentage of total variance accounted for by rotated factors.

Note. Bold typeface indicates factor loadings > .4. APM-SF is Raven’s Advance Progressive Matrices Short Form.
(Carroll, 1993; Schmid & Leiman, 1957). The two factor EFA solution was subjected to a Schmid-Leiman transformation using syntax developed by Wolff and Preising (2005). The obtained solution is presented in Table 1.4. It can be seen that the higher order general factor explained more than one quarter of the total variance and all of the variables have substantial loadings on this factor. It can also be seen that the first order factors are more clearly defined after the general factor loadings are taken into account. A first order WM/Gf factor was clearly loaded by Swaps, Dot Matrix and APM-SF and a first order Gs factor was clearly loaded by the paper-and-pencil speed tests. Space Code was substantially loaded on the general factor but had relatively small and almost equal loadings on both first order factors.

**Composite Ability Score Analyses**

To assess the relationship between the computer-game-like tests and the traditional ability tests more directly, composite Gs and WM/Gf scores were derived and correlated with the computer-game-like tests. Gs composites were derived by computing the average $z$ scores of Digit Symbol, Visual Matching and Decision Speed scores. A WM/Gf composite was formed using the Picture Swaps, Dot Matrix and APM-SF $z$ scores. Space Code was somewhat similarly correlated with both the Gs ($r = .52, p < .001$) and WM/Gf ($r = .47, p < .001$) composites. Space Matrix was very strongly correlated with the Wm/Gf composite ($r = .69, p < .001$) but more moderately with the Gs composite ($r = .40, p < .001$). The Gs and WM/Gf composite measures were correlated to the same extent as their respective EFA factors ($r = .45, p < .01$).
Table 1.4

**Summary of Schmid-Leiman hierarchial solution.**

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>WM/Gf</th>
<th>Gs</th>
<th>(h^2)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digit Symbol</td>
<td>.532</td>
<td>.086</td>
<td>.452</td>
<td>.495</td>
<td></td>
</tr>
<tr>
<td>Visual Matching</td>
<td>.588</td>
<td>.048</td>
<td>.547</td>
<td>.648</td>
<td></td>
</tr>
<tr>
<td>Decision Speed</td>
<td>.497</td>
<td>-.086</td>
<td>.589</td>
<td>.601</td>
<td></td>
</tr>
<tr>
<td>Picture Swaps</td>
<td>.532</td>
<td>.537</td>
<td>.001</td>
<td>.572</td>
<td></td>
</tr>
<tr>
<td>Dot Matrix</td>
<td>.596</td>
<td>.412</td>
<td>.191</td>
<td>.562</td>
<td></td>
</tr>
<tr>
<td>APM-SF</td>
<td>.484</td>
<td>.595</td>
<td>-.105</td>
<td>.600</td>
<td></td>
</tr>
<tr>
<td>Space Code</td>
<td>.522</td>
<td>.267</td>
<td>.262</td>
<td>.412</td>
<td></td>
</tr>
<tr>
<td>Space Matrix</td>
<td>.579</td>
<td>.509</td>
<td>.077</td>
<td>.601</td>
<td></td>
</tr>
<tr>
<td>Eigenvalue</td>
<td>2.36</td>
<td>1.16</td>
<td>0.97</td>
<td>4.49</td>
<td></td>
</tr>
<tr>
<td>% of Variance (^a)</td>
<td>29.5</td>
<td>14.5</td>
<td>12.1</td>
<td>56.1</td>
<td></td>
</tr>
</tbody>
</table>

\(^a\) Percentage of total variance accounted for.

Note. Bold typeface indicates factor loadings > .4. APM-SF is Raven’s Advance Progressive Matrices Short Form.

**Correlational Analyses of Non-Ability and Ability Measures**

The relationships that demographic and gaming measures had with ability composite scores and the game-like computerised tests were examined using correlations. Inspection of the correlations between gaming measures indicated that items 1 (On average how many hours do you play computer games each week) and 2 (On average how many hours do you play computer games using a mouse each week) were very strongly correlated \((r = .92, p < .001)\) while item 3 (At any stage in the past what was the maximum number of hours you played computer games in a week) was more moderately correlated with these, \(r = .45, p < .001\) and \(r = .26, p < .05\), respectively. Items 1 and 2 were combined to create a
‘Game Play’ measure and item 3 was retained as a separate measure, ‘Past Game Play’. Gender was significantly correlated with Game Play ($r = .47, p < .001$) and Past Game Play ($r = .35, p < .001$), the positive correlation indicating males had more game experience. Game experience measures were not significantly correlated with either age or handedness. Correlations between these measures and ability measures are presented in Table 1.5. It can be seen that gender was moderately correlated with both WM/Gf and Space Code. Game Play was moderately correlated with all ability measures except Gs while Past Game Play was moderately correlated with WM/Gf and somewhat more strongly with Space Code. To assess whether the relationship between gender and the computerised speed tests was mediated by gaming experience, partial correlations were obtained. After controlling for Game Play and Past Game Play the correlation between gender and Space Code was non significant and almost zero ($r = .06, p = .65$).
Table 1.5

Correlations of non-ability measure with ability composite scores and computer-game-like tests.

<table>
<thead>
<tr>
<th></th>
<th>Gs</th>
<th>WM/Gf</th>
<th>Space Code</th>
<th>Space Matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>.00</td>
<td>-.04</td>
<td>-.05</td>
<td>.02</td>
</tr>
<tr>
<td>Gender</td>
<td>-.15</td>
<td>.26*</td>
<td>.25*</td>
<td>.20</td>
</tr>
<tr>
<td>Handedness</td>
<td>.08</td>
<td>.03</td>
<td>-.12</td>
<td>-.12</td>
</tr>
<tr>
<td>Game Play</td>
<td>.05</td>
<td>.30*</td>
<td>.31*</td>
<td>.29*</td>
</tr>
<tr>
<td>Past Game Play</td>
<td>-.04</td>
<td>.25*</td>
<td>.42**</td>
<td>.10</td>
</tr>
</tbody>
</table>

* $p < .05$, ** $p < .01$.

Gender was coded such that positive correlations would indicate higher scores for males, negative correlations would indicate higher scores for females.

Handedness was coded such that positive correlations would indicate higher scores for right handed participants.

Study 2 – Rationale

In Study 1, and previous studies (McPherson & Burns, 2005, 2007), we have assessed the viability of using computer-game-like tests of cognitive abilities within university samples. However, we have also suggested that computer-game-like tests may be particularly suited for testing younger children. Indeed, the initial impetus for our research program came from our own frequent observations of boredom and motivational fatigue in school aged children undertaking more traditional assessments via computer. In Study 2 we thus aimed to assess the concurrent validity of Space Code and Space Matrix within a school aged sample. Additionally, we also sought to make a preliminary assessment of the external validity of the two computer-game-like tests.
In their recent large scale study, Luo, Thompson and Detterman (2006) demonstrated that Gs and WM measures predicted academic achievement as well as traditional Gc and Gf measures. Perhaps this is not altogether surprising considering the substantial amount of research indicating that Gs and WM measures share substantial variance with broader composite cognitive ability and/or Gf itself (see McGrew, 2005, for a review). However, it is important to note that a substantial portion of the relationship between Gs and more general ability is often indirect via its relationship to short term memory (Gsm) and WM ability. Additionally, it is possible that Gs and WM measures will differ in their relationships to specific areas of achievement (Floyd, Shaver, & McGrew, 2003).

Thus, in addition to assessing concurrent validity we aimed to assess the external validity of Space Code and Space Matrix as predictors of scholastic achievement. Furthermore, in order to address the possibility that Gs and WM measures may be differentially related to various achievement domains we used three external criterion measures; English, Mathematics and Science subject grades. Although specific predictions were difficult to make, based on results from the Woodcock Johnson III normative sample reported by Floyd, Shaver and McGew (2003), we expected Gs measures to better predict English grades than they would Mathematics or Science grades and WM measures, to the extent that they should share predictive power with Gf measures, would better predict Mathematics grades than they would English or Science grades. Furthermore, the stronger direct relationship between WM and more general ability suggested that WM measures may be stronger predictors than Gs measures in general. However, the primary aim was to directly compare the external validity of the computer-game-like tests and established ability marker tests.
Study 2 – Method

Participants

There were $N = 94$ participants who were children aged 11 to 14 years (Mean age = 12.88, $SD = 0.80$) recruited from 6th, 7th and 8th grades at a private school in Adelaide, South Australia. There were 42 males (Mean age = 12.73 years, $SD = 0.87$) and 52 were female (Mean age = 13.00 years, $SD = 0.73$).

Materials

Non-Ability Measures

Measures were the same as described in Study 1 except for age which was measured in years and months.

Tests of Processing Speed (Gs)

Digit Symbol. As described in Study 1.

Visual Matching. As described in Study 1.

Decision Speed. As described in Study 1.

Tests of Working Memory/ Fluid Intelligence (WM/Gf)

Picture Swaps. This test was the same as described in Study 1 with alterations made to accommodate the younger sample and stricter time limits. There were 24 items in total and four levels of difficulty; items 1-6 required one swap, items 7-12 required two swaps, items 13-18 required three swaps, and items 19-24 required four swaps. There was a 10 minute time limit implemented within the computer software. The dependent variable was the total number of correct responses completed within the time limit. Cronbach’s alpha was .87.

Raven’s Standard Progressive Matrices Short Form (SPM-SF). This was a computerised 32 item version of the original test (Raven et al., 1998). The 32
items used were the same subset used by Zajac and Burns (in press) with items selected on the basis of prior research investigating the factor structure of this test (Lynn, Allik, & Irwing, 2004). Items were presented and responded in the same manner as the APM-SF described in Study 1. There was a 10 minute time limit implemented within the computer software. The dependent variable was the total number of correct responses completed within the time limit. Cronbach’s alpha was .69.

_Counter-Game-Like Tests_

**Space Code.** This was the same test as described in Study 1. As in Study 1 the number of spaceships correctly destroyed in the first and second sixty second blocks of the test were recorded and correlated to provide a measure of split-half reliability \( r = .77, p < .001 \).

**Space Matrix.** This was the same test as described in Study 1 with the following alterations made to accommodate the younger sample. There were two practice ‘runs’ with one sectors displayed and one practice run with two sectors displayed. The test consisted of 30 sector locations that were presented in sequences of two and three sectors; there were six sequences of two sectors and six sequences of three sectors. The test took approximately 4 minutes to complete. Cronbach’s alpha was .70.

_Academic Achievement_

**Preparatory School Grades.** For students in 6th and 7th grades (n = 44) first semester school subject scores were obtained for English, Mathematics and Science. These grades were in the form of ordinal categorical grades (e.g., A, B, or C). These grades were recoded as minimum percentages required for each grade as follows, A = 85, B = 70, C = 55. No student obtained a grade lower than a C.
High School Grades. For students in 8th grade (n = 50), first semester school subject scores were obtained for English, Mathematics and Science. These grades were raw percentages.

Procedure

Teachers distributed information sheets and parental/guardian consent forms to children within class. Students were asked to read the information sheet with their parent or guardian and return the signed consent form if they were interested in participating. The information sheet stressed that participation was voluntary, was not associated with school assessment in any way, and that results would be anonymous and confidential.

Participants were tested in groups of two to four in a quiet room and asked to concentrate and do their best for each individual test. Tests were administered to all participants in the following order: Non-Ability Measures; Space Code; Space Matrix, Visual Matching; Decision Speed; Digit Symbol; Swaps; RSPM-SF. A consistent test order was used in an effort to make any fatigue or motivational effects equal across all participants. Sessions took approximately 45 minutes to complete.

All computerized tests were run on a Pentium 4 class computer (2.4 MHz) with a Microsoft optical Intellimouse. Display was via a 19 inch LCD monitor with a vertical refresh rate of 100 Hz and a screen resolution of 1280 x 1024 pixels.

Study 2 – Results

Descriptive statistics for all tests are presented in Table 2.1. Both the Picture Swaps and SPM-SF had means of about half the possible maximum scores (24 and 32 respectively). The mean for Space Matrix was similarly slightly below half
Table 2.1

Descriptive statistics for ability and computer-game-like tests.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digit Symbol</td>
<td>60.3</td>
<td>12.1</td>
<td>31</td>
<td>110</td>
</tr>
<tr>
<td>Visual Matching</td>
<td>520.2</td>
<td>20.8</td>
<td>473</td>
<td>564</td>
</tr>
<tr>
<td>Decision Speed</td>
<td>517.0</td>
<td>15.3</td>
<td>480</td>
<td>543</td>
</tr>
<tr>
<td>Picture Swaps</td>
<td>11.45</td>
<td>5.0</td>
<td>1</td>
<td>24</td>
</tr>
<tr>
<td>SPM-SF</td>
<td>16.3</td>
<td>4.0</td>
<td>8</td>
<td>25</td>
</tr>
<tr>
<td>Space Code</td>
<td>49.1</td>
<td>8.7</td>
<td>20</td>
<td>73</td>
</tr>
<tr>
<td>Space Matrix</td>
<td>13.1</td>
<td>4.6</td>
<td>4</td>
<td>27</td>
</tr>
<tr>
<td>Space Matrix Speed</td>
<td>77.2</td>
<td>22.6</td>
<td>87</td>
<td>171</td>
</tr>
</tbody>
</table>

a Based on Rasch scaled W scores centred on 500.

Note. SPM-SF is Raven’s Standard Progressive Matrices Short Form. Space Matrix Speed is the number of ships destroyed in Space Matrix.

of the possible maximum score (30). An independent samples t test indicated the Space Code mean for the Study 1 sample ($M = 61.0, SD = 9.1$) was higher than for the present sample, $t(162) = 8.50, p < .001, d = 1.34$. The Study 1 sample also had developmental increases in Gs across these age groups (Horn & Noll, 1997). The smaller effect size for Space Code suggests the task was not particularly difficult, relative to the other Gs tasks, for the younger sample. Comparisons for other tests could not be made because these were not equivalent for both samples.

Test Intercorrelations

Correlations between all ability measures are presented in Table 2.2. Relationships between the Gs and WM/Gf marker tests were generally as
theoretically expected, although, the magnitude of correlations was generally smaller than in Study 1. The three Gs marker tests, Digit Symbol, Visual Matching, and Decision Speed were all significantly correlated with each other as were the two WM/Gf tests, Picture Swaps and SPM-SF. Picture Swaps also showed substantial relationships with two Gs marker tests, Digit Symbol and Visual Matching. Space Code was only significantly correlated with Digit Symbol and SPM-SF while Space Matrix was significantly correlated with every other test. Correlations between Space Matrix and ability tests were of a similar magnitude to the correlations between the marker tests themselves, while for Space Code these correlations were smaller. As in Study 1, Space Matrix Speed was strongly related to Space Code but not with the primary Space Matrix memory measure, once again indicating independence of the speed and working memory components of the test.
Table 2.2

*Correlation matrix for ability and computer-game-like tests.*

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Digit Symbol</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Visual Matching</td>
<td>.57**</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Decision Speed</td>
<td>.42**</td>
<td>.40**</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Picture Swaps</td>
<td>.32**</td>
<td>.33**</td>
<td>.14</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>SPM-SF</td>
<td>.23†</td>
<td>.15</td>
<td>.19</td>
<td>.38**</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Space Code</td>
<td>.29**</td>
<td>.16</td>
<td>.20</td>
<td>.20</td>
<td>.36**</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Space Matrix</td>
<td>.31**</td>
<td>.23†</td>
<td>.22†</td>
<td>.35**</td>
<td>.44**</td>
<td>.34**</td>
<td>-</td>
</tr>
<tr>
<td>8</td>
<td>Space Matrix</td>
<td>.03</td>
<td>-.03</td>
<td>-.11</td>
<td>.15</td>
<td>.24†</td>
<td>.61**</td>
<td>.05</td>
</tr>
</tbody>
</table>

Speed

* p < .05, ** p < .01.

Note. SPM-SF is Raven’s Standard Progressive Matrices Short Form.
Exploratory Factor Analyses (EFA)

The tests included in EFA procedures were the three Gs markers (Digit Symbol, Visual Matching and Decision Speed), the two WM/Gf measures (Picture Swaps and SPM-SF), and the two computer-game-like tests, Space Code and Space Matrix. As with the APM-SF in Study 1, the SPM-SF was expected to load with the WM tasks in the absence of other Gf marker tests. The Kaiser-Meyer-Olkin (.764) and Bartlett ($\chi^2 = 127.59, \text{df} = 21, p < .001$) statistics indicated the correlation matrix was suitable for EFA. An initial Principal Components Analysis resulted in two components with eigenvalues larger than unity (2.80 and 1.21 respectively) and a third eigenvalue of 0.85. The first two components accounted for 40.0% and 17.4% of the total variance respectively. As for Study 1, subsequent analyses were conducted using Principal Axis Factoring (PAF).

The number of factors to extract was decided using the same criteria as in Study 1. Visual inspection of the scree plot and Parallel Roots Analysis suggested a two factor solution while Velicer’s Minimum Average Partial test suggested a one factor solution. A one factor PAF solution was judged unsuitable because there were ten residual correlations (47.6%) larger than .10. A two factor PAF solution was next obtained explaining 43.1% of the common variance. The two factors were then rotated using the oblique Direct Oblimin procedure (Delta = 0). The correlation between the two factors was -.50. There were no residual correlations larger than .10. Inspection of the Pattern Matrix (Table 2.3) indicated two quite distinct factors. Factor 1 was strongly loaded by SPM-SF and Space Matrix, and, to a lesser degree Picture Swaps and Space Code. Factor 2 was strongly loaded by Digit Symbol and Visual Matching with Decision Speed also having a smaller but substantial loading. Considering the Structure Matrix it can be seen that all of
### Table 2.3

**Summary of two factor PAF solution.**

<table>
<thead>
<tr>
<th></th>
<th>Pattern Matrix</th>
<th></th>
<th>Structure Matrix</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Factor 1</td>
<td>Factor 2</td>
<td>Factor 1</td>
<td>Factor 2</td>
<td>Factor 2</td>
<td>h^2</td>
</tr>
<tr>
<td>Digit Symbol</td>
<td>.081</td>
<td>-.729</td>
<td>.445</td>
<td>-.769</td>
<td></td>
<td>0.597</td>
</tr>
<tr>
<td>Visual Matching</td>
<td>-.081</td>
<td>-.813</td>
<td>.325</td>
<td>-.772</td>
<td></td>
<td>0.654</td>
</tr>
<tr>
<td>Decision Speed</td>
<td>.061</td>
<td>-.487</td>
<td>.304</td>
<td>-.518</td>
<td></td>
<td>0.271</td>
</tr>
<tr>
<td>Picture Swaps</td>
<td>.418</td>
<td>-.178</td>
<td>.507</td>
<td>-.387</td>
<td></td>
<td>0.281</td>
</tr>
<tr>
<td>SPM-SF</td>
<td>.811</td>
<td>.151</td>
<td>.735</td>
<td>-.254</td>
<td></td>
<td>0.634</td>
</tr>
<tr>
<td>Space Code</td>
<td>.463</td>
<td>-.064</td>
<td>.496</td>
<td>-.296</td>
<td></td>
<td>0.249</td>
</tr>
<tr>
<td>Space Matrix</td>
<td>.615</td>
<td>-.056</td>
<td>.643</td>
<td>-.363</td>
<td></td>
<td>0.416</td>
</tr>
<tr>
<td>Eigenvalue^a</td>
<td>1.85</td>
<td>1.89</td>
<td>1.85</td>
<td>1.89</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Cumulative %)</td>
<td></td>
<td></td>
<td>(41.30)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

^a Eigenvalues presented are for extracted factors after rotation.

^b Indicates percentage of total variance accounted for by rotated factors.

Note. Bold typeface indicates factor loadings > .4. SPM-SF is Raven’s Standard Progressive Matrices Short Form.
the tests are at least moderately correlated with both factors but the pattern of relationships is similar to the Pattern Matrix. Factor 1 is most strongly loaded by SPM-SF, Space Matrix and Picture Swaps and is thus interpreted as a WM/Gf factor. Factor 2 is most strongly loaded by Visual Matching, Digit Symbol and Decision Speed and is thus interpreted as a Gs factor. These results further support the concurrent validity of Space Matrix as a measure of WM/Gf. However, with Space Code loading more strongly on the WM/Gf in both the Pattern and Structure Matrix, the concurrent validity of Space Code as a Gs measure is not supported.

Composite Ability Score Analyses

To assess the relationship between the computer-game-like tests and the traditional ability tests more directly, composite Gs and WM/Gf scores were derived and correlated with the computer-game-like tests. Gs composites were derived by computing average $z$ scores for Digit Symbol, Visual Matching and Decision Speed scores. A WM/Gf composite was formed using the Picture Swaps and SPM-SF $z$ scores. Space Code was more strongly correlated with the Wm/Gf composites ($r = .34, p < .001$) than with the Gs composite ($r = .27, p < .01$). Space Matrix was quite strongly correlated with the Wm/Gf composite ($r = .48, p < .001$) and moderately with the Gs composite ($r = .31, p < .001$). The Gs and WM/Gf composite measures were less strongly correlated than the respective EFA factors ($r = .34, p < .01$).

Correlational Analyses of Non-Ability and Ability Measures

The relationships that demographic and gaming measures had with ability composite scores and the game-like computerised tests were examined using correlations. Inspection of the correlations between gaming measures indicated that Items 1 (On average how many hours do you play computer games each
week) and 2 (On average how many hours do you play computer games using a mouse each week) were strongly correlated ($r = .73, p < .001$) while Item 3 (At any stage in the past what was the maximum number of hours you played computer games in a week) was strongly correlated with Item 1 ($r = .69, p < .001$) but more moderately with Item 2 ($r = .42, p < .001$). To make results consistent with Study 1, Items 1 and 2 were combined to create a ‘Game Play’ measure and item 3 was retained as a separate measure, ‘Past Game Play’. Gender was significantly correlated with Game Play ($r = .38, p < .001$) and Past Game Play ($r = .25, p < .05$), the positive correlation indicating males had more game experience. Game experience measures were not significantly correlated with either age or handedness. Correlations between the non-ability measures and ability measures are presented in Table 2.4. In line with well documented developmental trends within these age groups (Horn & Noll, 1997) age was significantly correlated with Gs, WM/Gf, and Space Code. However, there was no significant relationship with Space Matrix. Gender was not significantly correlated with any of the ability measures. Game Play was correlated with WM/Gf and Space Code.
Table 2.4

*Correlations of non-ability measures with ability composite scores and computer-game-like tests.*

<table>
<thead>
<tr>
<th></th>
<th>Gs</th>
<th>WM/Gf</th>
<th>Space Code</th>
<th>Space Matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>.33**</td>
<td>.40**</td>
<td>.26*</td>
<td>.14</td>
</tr>
<tr>
<td>Gender a</td>
<td>-.18</td>
<td>.06</td>
<td>.18</td>
<td>.14</td>
</tr>
<tr>
<td>Handedness b</td>
<td>.14</td>
<td>.01</td>
<td>.07</td>
<td>-.10</td>
</tr>
<tr>
<td>Game Play</td>
<td>.04</td>
<td>.22*</td>
<td>.25*</td>
<td>.14</td>
</tr>
<tr>
<td>Past Game Play</td>
<td>-.06</td>
<td>.15</td>
<td>.17</td>
<td>.20</td>
</tr>
</tbody>
</table>

* p < .05, ** p < .01.

a Gender was coded such that positive correlations would indicate higher scores for males, negative correlations would indicate higher scores for females.

b Handedness was coded such that positive correlations would indicate higher scores for right handed participants.

*External Validity Comparisons*

To assess the external predictive validity of the computer-game-like tests correlations were obtained between these tasks and school grades. Additionally, traditional marker tests for Gs and WM/Gf were chosen to provide comparative predictors. Digit Symbol was chosen as a comparative task for Space Code because it is a well known Gs task from what is perhaps the most widely known and used ‘IQ’ battery, the Wechsler Scales, and both Space Code and Digit Symbol are coding tasks. The SPM-SF was chosen as a comparative task for Space Matrix as it is perhaps the most widely recognized test of Gf and is also thought to be one of the best indicators of general cognitive ability (Carpenter et al., 1990; Fry & Hale, 1996; Jensen, 1998; Prokosch et al., 2005). Kendall’s tau correlations were used for assessing relationships with preparatory subject grades.
because these were categorical and there were numerous tied ranks (Howell, 2007).

Considering Digit Symbol and Space Code it can be seen that Space Code showed no significant relationships with school subject grades in either sample. Digit Symbol had a moderate significant relationship with Preparatory English grades. Considering SPM-SF and Space Matrix it can be seen that the WM tasks were generally better predictors of school grades than the speed tasks. SPM-SF and Space Matrix were significant predictors of preparatory English grades but not high school English. However, the only other significant correlation was between Space Matrix and mathematics grades in both samples. Correlations between SPM-SF and mathematics grades were non significant but only slightly smaller in magnitude. Space Matrix had the strongest correlations with all subject grades except for high school English and Science.
Table 2.5

Correlations for comparative ability and computer-game-like tests with school subject grades.

<table>
<thead>
<tr>
<th></th>
<th>Preparatory School Grades&lt;sup&gt;a&lt;/sup&gt;</th>
<th>High School Grades</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>English</td>
<td>Mathematics</td>
</tr>
<tr>
<td>Digit Symbol</td>
<td>.25&lt;sup&gt;*&lt;/sup&gt;</td>
<td>.07</td>
</tr>
<tr>
<td>Space Code</td>
<td>-.07</td>
<td>-.03</td>
</tr>
<tr>
<td>SPM-SF</td>
<td>.28&lt;sup&gt;*&lt;/sup&gt;</td>
<td>.22</td>
</tr>
<tr>
<td>Space Matrix</td>
<td>.35&lt;sup&gt;**&lt;/sup&gt;</td>
<td>.32&lt;sup&gt;*&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

<sup>a</sup>Correlations for Preparatory School Grades are Kendall’s tau nonparametric correlation.
General Discussion

The results of both studies provide preliminary support for the potential of developing computer-game-like tests of information processing abilities. In both samples, Space Code and Space Matrix were found to have acceptable internal reliabilities, and, within the university sample, both tests provided stable results across two administrations. The computer-game-like tests also shared substantial variance with the traditional abilities tests. However, there were differential levels of support for each test’s concurrent and external validity.

In Study 1, Space Code loaded, relatively weakly, on both a Gs and a WM/Gf factor in the initial EFA solution. In a subsequent Schmid-Leiman transformation, Space Code had a substantial loading on the higher order factor but relatively weaker loadings on the first order Gs and WM/Gf factors. Thus, although Space Code appeared to share considerable variance with the traditional tests, it did not exhibit discriminant validity as a measure of Gs. In Study 2, Space Code actually loaded more strongly on a WM/Gf factor. In an earlier paper (McPherson & Burns, 2007) we reported that a simple prototype version of Space Code loaded on a Gs and a visuo-spatial ability (Gv) factor. However, in a second study incorporating a larger battery of tests, the final version of Space Code was primarily loaded on a Gs factor. Overall, these results suggest Space Code is probably a mixed ability measure and will require further refinement if it is to function as a measure of Gs.

Space Code was originally based on a speeded computerised coding task, Symbol Digit, that utilised simple symbols similar those used in the well known Digit Symbol subtest from the various Wechsler scales (McPherson & Burns, 2005). The Symbol Digit test was found to load saliently on a Gs factor across three separate studies. Space Code and Symbol Digit use the same presentation and response methods but Space Code uses multi-coloured spaceships of various
designs as coding symbols. For this reason, we suggest that the more complex loading patterns observed for Space Code are probably due to the more visually complex stimuli used. As pointed out by a reviewer of the current paper, the use of more complex visual stimuli would also make greater demands on visual memory which may also explain why the task had considerable loadings on the WM/Gf factor. We believe the mixed loadings are less likely to be due to the inclusion of dynamic game-like elements, such as scoring and auditory feedback, because loadings on a Gv factor were observed for a prototype version that did not include these elements (McPherson & Burns, 2007).

In contrast to Space Code, Space Matrix exhibited quite consistent psychometric properties across both studies. In Study 1 and Study 2, Space Matrix exhibited a pattern of loadings similar to the traditional WM and Gf measures. Although WM and Gf are distinct constructs there also exists a substantial amount of shared variance at the latent construct level (Ackerman et al., 2005; Engle et al., 1999; Fry & Hale, 2000; McGrew, 2005; Unsworth & Engle, 2005). Thus, in both of the present studies, in the absence of additional Gf measures, the Raven’s, a measure of Gf, and WM measures loaded on the same factor. Space Matrix also had substantial relationships with the Raven’s tests in both studies (.53 and .44, respectively). Previous researchers have reported correlations of about .30 - .35 between well established WM measures (e.g, reading, counting and operations span) and Raven’s measures (Conway et al., 2002; Engle et al., 1999; Unsworth & Engle, 2005). Thus, Space Matrix would seem to compare quite favourably as a correlate of Gf, or even g, for those who argue that the Raven’s are perhaps the best measures of general intelligence (Jensen, 1998; Prokosch et al., 2005). These results were further supported by comparisons of respective relationships with school grades.
Space Matrix and the Raven’s tests had a similar pattern of relationships with school subject grades in both samples and the magnitude of the correlations between Raven’s and school grades were similar to those reported by other researchers (Rushton, Skuy, & Fridjhon, 2003). Although the present samples are too small to form the basis for any strong conclusions, it was encouraging to find that correlations between school grades and Space Matrix were higher than those with the Raven’s tests for all subjects except High School Science. Interestingly, both tests had stronger relationships with Preparatory English than High School English which may reflect curriculum and/or assessment differences or developmental trends. However, this is difficult to ascertain without detailed knowledge of course components and assessment procedures. The significant relationships Space Matrix had with Preparatory English and both Preparatory and High School Mathematics is consistent with research indicating that WM plays a role in reading and writing abilities (Floyd et al., 2003; Hitch, Towse, & Hutton, 2001; Just & Carpenter, 1992) and in mathematical abilities (Floyd et al., 2003; Hitch et al., 2001; Swanson & Kim, 2007b). In contrast, Space Code had no substantial relationships with any of the school subject grades. Digit Symbol, however, was significantly related to Preparatory English and slightly less, albeit non significantly, with High School English, which is consistent with previous research indicating a relationship, at this age level, between Gs and basic reading skills (Floyd et al., 2003). On the whole, these results provide encouraging preliminary evidence for the external validity of Space Matrix but not for Space Code. Space Code, in particular, will require further development before it can be used as an alternative measure of processing speed. Current versions of Space Code have been designed to reduce the visual and working memory demands of the task, in an effort to redress the observed relationships with these constructs.
Concerning the non ability measures there were a number of interesting findings in both studies. In Study 1, there was a significant positive correlation between gender and Space Code indicating that males were quicker in this test. However, after controlling for the frequency of current and past computer gaming, the correlation was reduced to almost zero. Thus, it appeared that the relationship was mediated by levels of computer gaming experience. Computer gaming measures had substantial positive relationships with Space Code and Space Matrix but they were similarly related to the WM/Gf composite. In Study 2, gender was not significantly correlated with either computer-game-like test but current computer gaming was significantly correlated with Space Code scores. However, as in Study 1 this measure was also similarly related to WM/Gf composites. Thus, although it could be argued that computer-game-like tests may be biased towards those who play more computer games, this would also seem to be the case for the WM/Gf measures. However, as the WM/Gf measures were also administered via computer, those who play more computer games may simply be more comfortable with computer administration of any test.

Future research may investigate whether this relationship will hold for WM and Gf measures administered via paper-and-pencil or oral methods. If the relationship is not dependent on administration mode a further possibility is that computer game playing may enhance attentional, visuospatial or other skills tapped by all of these tasks. Another possibility is that those with higher WM/Gf ability are more likely to enjoy playing computer games. Recent studies have begun to address some of these possibilities (Aliya, 2002; Castel et al., 2005; Enochsson et al., 2004; Quaiser-Pohl et al., 2006) but this issue is clearly beyond the scope of the present studies. However, it remains important to acknowledge that computer-game-like tests may not rank individuals in the same way as traditional tests even
where a certain degree of construct validity is achieved. Future research will need to address this issue in greater detail. A number of other issues will also need to be addressed.

The present studies involved small sample sizes and a limited range of ability levels and ages. Although the results are thus far encouraging, in particular for Space Matrix, research will need to be conducted using larger and more representative samples. Additionally, and perhaps more importantly, the range of tests used to assess construct validity needs to be expanded to represent a much larger range of cognitive abilities. In the absence of a wider range of tests it is not possible to ascertain whether these computer-game-like tests are not also measuring other abilities. Finally, another more pragmatic issue that arose would seem to highlight the many potentially interesting pitfalls that may lie ahead.

In line with research indicating game players commonly adopt a trial and error approach when learning a new game (Miller, Schweingruber, & Brandenburg, 2001; Mitchell & Savill-Smith, 2004) some children, in both preparatory and high school samples, did not stop to read the instructions before beginning the actual test. Future versions of the computer-game-like tests will address this issue by incorporating more extensive trial and error practice. The challenge will be to maintain as much of the gaming experience as possible, with all of the potential advantages, while also attempting to make tools that are as reliable and valid as traditional measures.
Chapter 6. General Conclusion

Overview of Thesis Aims

The broad aim of the present thesis was to develop computer-game-like tests of processing speed (Gs) and working memory (WM). This aim met with some successes and some failures. Two computer-game-like tests (Space Code and Space Matrix) were developed and assessed across a series of five studies. In addition to these two tests, a small number of simpler mouse response speed tests were developed. The validity of the computer-game-like tests was assessed using a range of established marker tests and, in the final study, school grades. Before making some general conclusions, a brief review of each study is provided together with any additional contextual conclusions not presented in the original papers.

Review of Studies

Study 1

The primary aim of Study 1 was to assess whether the mouse response method developed would be appropriate for the measurement of Gs. It was also important to assess whether the response method might also share variance with a speed construct (Gt) measured by so-called elementary cognitive tasks. These two related aims were assessed via two tests using the mouse response method and marker tests for both Gs and Gt. The Gs tests were paper-and-pencil tests while the Gt tests used a custom apparatus controlled by computer (i.e., a Jensen Box). The results suggested that the two mouse response tests were tapping Gs rather than Gt.

These findings were in line with those obtained by Parks et al. (2001) who developed a Gs search task using a mouse response method with similar results. However, the battery of tests used in Study 1 was very small and the factors may
reflect unique variance associated with the specific type of Gs and Gt tasks used. Whatever the case may be, Study 1 indicated that the mouse response tests shared considerable variance with the paper-and-pencil Gs tasks rather than the Gt tasks.

**Study 2**

The aim of Study 2 was to expand on the test developed in Study 1 (Symbol Digit) by incorporating a computer-game-like cover story and associated graphics and sounds. This was done to assess whether the added visual and other complexities would interfere with the measurement of Gs, or whether this test would also tap visual ability (Gv), or both of them. The overall results of Study 2 indicated that the new test (Space Code) loaded both Gs and Gv tests. This result suggested the possibility that Space Code may be factorially complex due to the increased visual demand of the computer-game-like medium. It is not uncommon for established ability tests to be factorially complex (Carroll, 1993; McGrew & Flanagan, 1998) and so the observation that Space Code loaded on two factors was not viewed as catastrophic for the development process.

**Study 3**

Study 3 aimed to assess a more computer-game-like version of Space Code (Space Code Game). Space Code Game was developed by adding a wider range of more dynamic elements such as scoring, timing, bonus points and a greater degree of response feedback. A further aim of Study 3 was to provide a stronger test of validity by using a broader battery of ability marker tests, including a larger sample of Gs tests to better measure the targeted construct.

Results from Study 3 suggested that Space Code Game loaded strongly with all of the traditional Gs measures and also loaded on a Gv factor at about the same level as the Gs measures did. Space Code Game also showed good
reliability within the session. Both of the computerised coding tasks (Symbol Digit and Space Code Game) showed substantial relationships with a short form of the Raven’s Advanced Progressive Matrices, suggesting the type of positive manifold one would expect of psychometrically sound ability measures.

Study 4

Study 4 had three primary aims. The first was to assess the construct validity of the new game-like WM test, Space Matrix. The second was to again assess the validity of Space Code (in Space Code Game format). It was also important to assess these two tests together because Space Matrix included Space Code as an embedded competing distractor task. A third aim was to assess the relationship between performance on these tests and external non-ability measures such as gender and previous computer game experience.

Space Matrix showed good initial construct validity as a test of WM, loading clearly with the other WM tasks and the Raven’s short form. This result also agreed with the large body of evidence linking WM and Fluid Intelligence (Gf) measures; the Ravens test being perhaps the archetypal Gf measure. Space Matrix was also reliable across two administrations within the testing session.

In contrast, the Space Code format showed poorer validity evidence than it had in the previous studies. Space Code had only small-to-moderate loadings on both WM and Gs factors. A Schmid-Leiman transformation of the solution suggested that Space Code had a substantial loading on a factor common to all of the tests but with this variance extracted it did not share much variance with the paper-and-pencil Gs tests.

Considering the relationships with non-ability variables, it was found that gender had relationships with a number of tests, but most particularly, with Space Code. However, after controlling for computer game playing experience this
relationship was eliminated. Game playing measures were related to performance on computer-game-like tests but also to performance on the WM/Gf measures. Handedness was not related to any other measure.

**Study 5**

The final study within the thesis research program aimed to take the two computer-game-like tests into the ‘real world’ and to one of the most likely target sample groups (school children) that such tests might be used with in applied settings. The aim within this context was to assess the construct validity of the tests in such a sample and, furthermore, to compare the external validity of the tests against some more established ability tests. As in Study 4, relationships between ability and non-ability measures such as gender and computer game experience were also assessed.

In regards to construct validity and relationships with non-ability measures, the results from Study 5 were very similar to those from Study 4. Space Code exhibited cross loadings on both a Gs and a WM/Gf factor while Space Matrix exhibited strong construct validity as a measure of WM/Gf. As in Study 4, computer game experience measures were related to the computer-game-like tests but also to the WM/Gf composite measure. As expected, age had a moderate and positive relationship with ability measures and the computer-game-like tests.

The most important aspect of Study 5, relative to all the previous studies, was the availability of external outcomes that are generally known to be positively related to the ability constructs that were assessed. Once again the results were much more positive for Space Matrix than they were for Space Code. While Space Code was unrelated to any of the school grade outcomes, Space Matrix exhibited
a pattern of relationships with school grades that was largely in line with those obtained for the Raven’s Progressive Matrices.

**Summary of Main Findings**

Across five studies it was found that prototype computer-game-like tests of cognitive abilities had psychometric properties that were in many respects comparable to those of traditional tests. Perhaps this is not altogether surprising considering that the tests were to some extent structurally based on already established tests. The two endpoint computer-game-like tests developed, Space Code and Space Matrix, showed reliability across separate administrations (albeit within the same sessions) as well as substantial, although inconsistent for Space Code, levels of validity relative to a moderate range, but carefully selected set, of criterion tests. In addition to this it was found that a computer mouse response method could be successfully implemented to measure a Gs construct similar to that measured with paper-and-pencil speed tests. Indeed a recurrent finding was that even the simplest measures designed to measure speed using this response method were consistently related to speed of performing paper-and-pencil Gs tests.

As previously mentioned, validity evidence for the final two computer-game-like tests was considerably less favourable for Space Code than Space Matrix. Space Code showed mixed loadings in Study 2, Study 4 and Study 5. It seems likely that Space Code is not a pure measure of Gs but may instead tap a range of abilities. In Study 4, it also had a strong loading on a common factor made up of Gs and WM tasks. This could be taken to indicate that it may tap important elementary processing abilities. However, considering the minimal relationships it had with school performance in Study 5 this indication is somewhat incompatible with the
findings of Luo et al. (Luo et al., 2006) suggesting that such measures should be strong predictors of scholastic achievement. A possible explanation is that Space Code caused a substantially stronger cognitive load on younger children which acted to further dilute the construct validity of a test that already showed mixed construct validity in older samples. I believe Space Code will need to be simplified to more specifically measure Gs and to more consistently exhibit acceptable construct validity (see below).

In contrast, Space Matrix showed strong construct and predictive validity in a university sample and two school aged samples. Space Matrix had clearly defined loadings on WM/Gf factors, clear relationships with WM/Gf composite scores, and had similar predictive validity for school grades as the Raven’s test. Despite being a relatively complex task, children readily understood the requirements. Additionally, and in line with suggestions of Luo et al. (Luo et al., 2006) that processing measures are easily manipulable relative to traditional ability measures, I was able to effectively change the difficulty level of Space Matrix for this younger sample without any major redevelopment of items. Such manipulation may also prove useful in the further refinement of Space Code into the future.

**Overall Significance**

The present thesis represents the first attempt to develop computer-game-like assessments to measure well established human cognitive ability traits within contemporary CHC theory. Additionally, results obtained also contribute to our understanding of computerised assessment more broadly as it relates to traditional assessment methods.

The successful development of a computerised Gs test using a computer mouse response method represents a potential new direction for computerised
speed tests which have traditionally utilized keyboard response methods or
customized response devices. This may have particular import in applied settings
with individuals familiar with graphical mouse oriented interfaces and where
specialized equipment is not readily available or feasible. The present thesis also
draws attention to important aspects of designing mouse response formats such
as restricting cursor movement and providing strong visual feedback to the user to
increase usability. Furthermore, the present thesis also represents an important
investigation into the relationship between mouse dexterity and cognitive abilities.
The results suggest that the ability to rapidly use a computer mouse to interact
with a graphical interface may be underpinned by similar processes underlying
performance on paper-and-pencil Gs tasks.

As mentioned above, the present thesis represents a first major attempt to
develop computer-game-like assessment tools that map directly onto
contemporary CHC ability constructs. Previous research has instead relied upon
pre-existing computer games or developed laboratory tools without specific
attention to construct validity. In line with much current research into cognitive
processing constructs the present thesis also contributes to the range of
assessment methods that might be employed in this ever expanding research
program. Results suggest that computer-game-like assessments can be
successfully implemented and may provide researchers with a broader range of
assessment options within theoretical and applied research.

Although results from the present thesis suggest that Space Code was not as
successful as Space Matrix, overall these results indicate substantial potential for
independent researchers to develop their own computer-game-like assessments.
Although it has often been considered unfeasible or extremely difficult to develop
computer-game-like software tools (Washburn, 2003) the present thesis
demonstrates that modern multimedia programming environments have considerably reduced the barriers to successful implementation of such tools. The software developed for this program of research was implemented using commercially available software and technical support obtained via online communities. Clearly the scope for software development will only improve into the future.

In the final study within this thesis Space Code and Space Matrix tests were used to predict school grades. This represents a first and significant attempt to link computer-game-like test performance with external outcomes generally considered centrally relevant to cognitive ability assessment. The results were also significant in that Space Matrix proved to be as strong a predictor of scholastic performance as the widely used Raven's test. This finding highlights the successful development of a computer-game-like test of WM/Gf that can be administered in group settings with relatively young children. Space Matrix can be implemented using standard PC equipment and takes around five minutes to complete. With a growing interest in WM as an important theoretical link between measures of cognitive processing capacity and problem solving, such a test may prove useful and convenient as a research tool and as a useful screening tool in applied settings.

Problems To Address

The present thesis suffered from a range of common empirical problems as well as some more specific problems. The most obvious empirical problem was the predominant use of small relatively homogeneous university samples. This was largely unavoidable due to the nature of institutional research participation schemes. The availability of participants within semester schedules and for limited
time durations also greatly restricted the size of test batteries that were possible for the studies. This pressure was even stronger within the school based study where an upper time limit of about 50 minutes was imposed. For this reason it was decided to specifically focus on establishing construct validity against sets of strong marker tests. The specific focus of such studies resulted in a number of other potential issues not being investigated.

One of the underlying assumptions outlined in Chapter 1 for investigating computer-game-like tests is that such tests will be more motivating and potentially less anxiety provoking for a large number of people. This assumption is relatively common amongst researchers interested in using computer-game-like formats (Berger et al., 2000; Bliss et al., 1991; Donchin, 1995; Jackson III et al., 1993; Porter, 1995; Ryan, 1994; Washburn, 2003). However, I have not been able to find any empirical demonstrations of such effects. The present thesis is also unable to provide a demonstration of such effects, the only potential evidence being limited to qualitative statements made by participants. For example, many participants requested additional attempts at the computer-game-like tests but not for the traditional tests. Additionally, it was also my opinion that research questions concerning motivation and test anxiety would most effectively be tackled after the construct validity of computer-game-like tests were established.

One of the central problems for researchers interested in test anxiety arises from the lack of a baseline performance measure. For instance, in an effort to isolate the effects of test anxiety researchers have used ability measures as a covariate but these measures are themselves confounded with test anxiety (McDonald, 2001). It is logical, or at least highly probable, that most ability tests would be perceived as threatening by individuals prone to test anxiety. Partialling on the basis of such measures can lead to underestimating the effects of test
anxiety. For this reason McDonald (2001) identified the need for obtaining control measures of ability that are not contaminated by test anxiety as the most pressing problem to be addressed in this research area. Computer-game-like tests are one of the obvious potential solutions because they may not so readily be perceived as tests.

However, for valid comparisons to be made it must first be clear what a computer-game-like test actually measures. Consider the case of a researcher who finds that a computer game is rated as less anxiety provoking than traditional tests and that it predicts outcomes better than traditional tests within highly test anxious individuals. This researcher may conclude that the computer game has measured performance better for these individuals because test anxiety has been reduced. However, it could also be argued that the game actually measured something different from the traditional tests. To separate the effects of anxiety and the underlying constructs being measured we need to have alternative tests, such as computer-game-like tests, that have established construct validities. The primary aim of the present thesis was to ascertain how viable this might be. However, studies incorporating anxiety, motivational and personality assessments are of considerable importance for more comprehensively understanding and utilizing the potential of computer-game-like tests (see below).

**Potential for Computer-Game-Like Assessments**

Carefully designed computer-game-like assessments have the potential to integrate features of traditional assessments and computer games. Thus, there is the possibility of having assessments that have known reliability and well defined construct validity while also providing a platform that is more appealing and intrinsically motivating, especially to populations that are traditionally difficult to
assess. For example, clinical populations with attentional disorders or acquired brain injuries. As mentioned in relation to test anxiety, computer-game-like tests could also provide an alternative measure of constructs that could serve as a viable comparative measure for certain populations who might be distinctly unmotivated by traditional tasks. Researchers have already begun to use existing computer games with clinical populations with some success.

Houghton et al. (Houghton et al., 2004) successfully used Crash Bandicoot, a Sony™ Play Station game, in a comparative study with Attention Deficit Hyperactivity Disorder (ADHD) children and a matched normal sample. The authors found that ADHD children performed quicker than normal children when working memory demands were lower but took longer when these demands were higher. However, with another outcome measure, number of obstacles navigated, they found that normal children performed better under low working memory demands but performed comparatively under higher working memory demands. Similarly, although ADHD children have been shown to perform poorly on computerised laboratory based tasks (Rubia, Taylor, & Sergeant, 1999), it has been suggested that this may reflect motivational aspects of the tasks (Lezak, Howieson, Loring, Hannay, & Fischer, 2004).

Indeed, in a previous study Houghton and colleagues (Lawrence et al., 2002) found that some aspects of executive function were not significantly different in ADHD children when operationalised as outcome measures within a computer adventure game. Although such research is clearly beneficial for our understanding of clinical populations the use of computer games with unknown construct validity makes results more difficult to integrate theoretically with other research. The specific development of computer-game-like tasks that can be
shown to measure the same constructs as traditional clinical tests could provide a more powerful point of comparison in this type of research.

Another potential for computer-game-like assessments is suggested by reports in the popular press of companies using computer games as assessment tools for specific professional skills. A BBC radio report (Logan, 2004) interviewed employers who were using Game Boy™ and online poker skills as key indicators of potential ability in digital trading environments. The rapid multi-tasking skills required in computer games being viewed as directly relevant to fast paced trading environments that use multiple screens and mouse clicks to interact with the markets. Additionally, online poker abilities were seen by some employers as a good indicator of how well a potential employee can balance emotional involvement in financial transactions and the mathematical ability to keep track of cards. Such applications point to unique features of computer games that may touch upon important psychological questions.

The development of computer-game-like assessment tools that function similarly to popular computer games and online games such as poker hold the possibility of combining ability assessment and personality assessment. For instance if it is found that certain scenarios within a poker game or an adventure game will evoke differential behaviour from individuals with different ability or personality traits then computer-game-like assessments could exploit this by artificially over sampling these particular contexts to provide more systematic and comprehensive measurement of the underlying trait. For example, certain situations within an adventure game may be known to be particularly difficult and require rapid problem solving, or even require risk taking behaviours, with relatively few players successfully negotiating them in a first attempt. These situations could be programmed to occur more often and potentially provide
measurement reliability via multiple measurements of the underlying behaviour traits. Such assessments could potentially retain their familiar game-like characteristics while also integrating psychometric properties to aid the interpretability of research and applied outcomes.

**Future Directions**

A number of future directions are suggested by the results of the present thesis as well as recent research and popular trends as discussed above. The particular finding that Space Code exhibited somewhat inconsistent and complex construct validity suggests that the test should be further refined if it is to measure Gs more purely. Versions of the test are currently being developed with a series of variations to address the increased complexity of the task relative to traditional paper-and-pencil Gs tasks. These variations include removing the coding aspects of the task by requiring players to simply match stimuli to destroy ships; using stimuli that exhibit more structural definition (e.g. are based on clearly differentiated shape prototypes such as X-wings); and using fewer stimuli variants as targets. Such manipulations aim to reduce memory and visual demands and thus increase performance covariance with Gs measures.

Space Matrix, having shown promising psychometric properties, should now be assessed more comprehensively within a larger network of ability tests and with larger and more diverse samples. Additionally, questions concerning motivation and anxiety could now be approached using this test. Studies using Space Matrix and other traditional tests as well as measures of test anxiety, test motivation, computer anxiety, overall enjoyment and external outcome measures could begin to assess whether Space Matrix is experienced differently from traditional tests and whether or not this has resulted in improved measurement for certain
individuals. Comparisons could be made between performance on computer-game-like tests and traditional tests under conditions aimed at varying performance motivation. For example it could be tested whether performance on computer-game-like tests matches performance on traditional tests where external rewards are linked to performance. Such methods could begin to assess the role of motivation in test performance and potentially validate the assumption that computer-game-like assessments are more motivating for some individuals. Additionally, relationships with personality traits may also be assessed. These investigations may also involve more in-depth analyses of response variables such as error rates, improvement rates, peak performance and fatigue effects.

Research might also now focus on developing computer-game-like tasks of other known cognitive abilities. A similar approach to the current thesis could also be taken whereby established tests are used as the basis for new tests. Alternatively, another approach, mentioned previously, is to take already known computer game formats and create specific versions of these to more closely match traditional test requirements. Such tests might also be termed computer-game-like tests because they may lose elements of randomness present in true computer games but could provide greater measurement stability and thus reliable psychometric information.

Concluding Statement

The present thesis represents a first step towards developing computer-game-like assessment tools that can potentially be integrated with our current knowledge and theories of human cognitive ability. The scope of this development was modest but the level of success substantial enough to warrant some optimism. The approach taken was to develop new assessment tools but with a strong
emphasis on using the measurement ground already mapped to contextualize and guide this development. This is certainly not the only approach that might be taken to advance and broaden the scope of cognitive ability assessment but may provide a first step for a research program linking computer-game-like assessment with the tremendous amount of work already devoted to understanding the structure of human cognitive abilities.
References


References. Page 161


Frearson, W., & Eysenck, H. J. (1986). Intelligence, reaction-time (Rt) and a new odd-man-out Rt paradigm. *Personality and Individual Differences, 7*(6), 807-817.


Deficit/Hyperactivity Disorder during computer game play. *British Journal of Educational Technology, 35*, 21-34.


Behavior Research Methods, 37(1), 48-58.


Appendix A

Appendix A. Paper 1 Reprint

*Behavior Research Methods, v. 37 (3), pp. 538-544, August 2005*

NOTE: This publication is included in the print copy of the thesis held in the University of Adelaide Library.
Appendix B

Appendix B. Paper 2 Reprint


NOTE: This publication is included in the print copy of the thesis held in the University of Adelaide Library.
CD-Rom contains the range of custom software developed within the PhD research program and referenced in the published papers.