Corvus: An Automatic Raven's-like Test Generator

Isaac Thimbleby

St Catherine's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity, 2018

Abstract

The primary focus of this thesis is the development of an automatic Raven's-like test generator, called Corvus. Corvus was designed to be capable of investigating the relative validity of repeat testing for Raven's Standard Progressive Matrices (SPM). Raven's SPM is often considered a 'Gold Standard' for general intelligence testing. However, following studies conducted using Corvus, this thesis found evidence that Raven's SPM lacks validity as a measure of intelligence on repeat testing. Specifically that retesting with Raven's SPM resulted in the tests losing correlation between item difficulty and time taken per item, while maintaining participant rank order (Spearman's $\rho, n = 53$).

An additional study using Corvus showed that mouse tracking can enhance information gain in online cognitive testing without significant impact to test score or time. An alternative way of scoring digit span tests was also investigated, and found to address common criticisms of traditional scoring methods.

This DPhil was written in an epidemiological context where studies routinely conduct repeat cognitive testing over sustained periods of time to assess cognitive trajectory. However, despite this, there is little research into the validity of repeated cognitive tests.

Corvus was designed through an ongoing iterative process together with a detailed quantitative analysis of established Raven's-like tests. Corvus is made available under an open source licence, for other researchers to use or develop, as are other pieces of code developed as part of this thesis.

Contents

Chapter 0

Preface

0.1 Acknowledgements

The completion of this DPhil would not have been possible without the help of the following people, and I would like to take this opportunity to thank them;

My supervisor Professor John Gallacher, who took a chance on me initially, and has given me the freedom to take this thesis where I wanted, while providing me with the scientific guidance I needed, as well as proofreading this thesis multiple times.

My father Professor Harold Thimbleby, who has been an incredible role model and teacher throughout my life, as well as having been of great help throughout this DPhill with proofreading, and in particular for his advice to start writing the thesis as soon as possible.

Dr Sarah Bauermeister, for making herself available, for her scientific and academic insight, for her knowledge of testing, for her proofreading this thesis, and for setting me deadlines.

My thanks also goes to the TrueColours team at Oxford University who kindly provided server space and use of their TrueColours system, which handled data storage, security and participant registration for my online studies, and I would also like to thank Simon Bond in particular for suggesting the name Corvus.

For their insightful thoughts and questions, I would like to thank my other colleagues and all the people I have been able to enjoy meeting and discussing my thesis with.

Lastly, I would especially like to thank my wife for moving to Oxford with me, as well as for her continuous friendship, forbearance, support, for proofreading this thesis with her extensive knowledge of the English language, and for providing me with tasty snacks (primarily halloumi).

0.2 Comments

Throughout this thesis I use a mix of active and passive mood. This variety is in accordance with the discussion of scientific English in the definitive *English Language* by David Crystal (Crystal 1995), and follows the tradition seen in Sir Isaac Newton's work, as witnessed in his effective variety of active and passive in his ground-breaking *Philosophiæ Naturalis Principia Mathematica*, as well as in the work of more recent leading science writers, such as Richard Feynman.

Due to the sheer quantity of computer program code written for the completion of this DPhil, at over ten thousand lines of code consisting primarily of JavaScript, Mathematica and HTML, the vast majority of it has not been included in this thesis, as this would add more than 200 pages (depending on font size). Instead

selections of various pieces of the programming have been included throughout the thesis when relevant, principally from Corvus.

The full body of Corvus's code, along with code for the Reverse Digit Span test has been uploaded to `https://github.com/Thimbleby?tab=repositories`, where they are made available under open-source licences. The online versions are machine readable and can be used directly by other researchers who are confident in programming in JavaScript. More details on the uploaded code can be found in Appendix B.

Please see the synopsis (Chapter 8) for details on some planned features that are yet to be implemented at the time of submitting this thesis. Some older versions of Corvus, such as the exact versions used in various studies presented here, as well as the smaller tests and other pieces of code used in this thesis, can also be made available by the author on request.

0.3 Dissemination

At the point of submission, work contained in this thesis has been presented in talks and displayed on posters at various conferences including the International Association for Computerized Testing 2015, Alzheimer's Association International Conference 2017, Alzheimer's Research UK 2018, and Dementias Platform UK 2018.

Chapter 1

Introduction

1.1 Summary

The accurate and valid measurement of cognitive change is important in epidemiology for describing cognitive trajectory in population studies, for detecting early cognitive impairment and for providing outcomes for clinical trials in dementia treatments.

Cognition is complex, and a broad range of cognitive domains are affected by ageing. Fluid intelligence is one such cognitive domain, which is defined as 'the ability to see relations among objects or the ability to see patterns in a repeating series of items ... a potential ability to learn and solve problems.' (Martin, Carlson & Buskist 2010).

This thesis focuses on fluid intelligence for a variety of reasons:

- Fluid intelligence is a cognitive domain of specific interest in epidemiological contexts.

- It is assessed via novel problems (Jaeggi, Buschkuehl, Jonides & Perrig 2008), and its validity directly interacts with retesting as problems are no longer novel on repeat assessment.

- On a practical level fluid intelligence tests are relatively easily systematised, which assists with the development of automatic test generators.

Fluid intelligence is well understood at baseline, and is the most researched non-verbal cognitive domain (McCallum, Bracken & Wasserman 2000). Yet relatively little research has been done on repeated assessment with fluid intelligence, though the same statement could be made of all cognitive domains (Scharfen, Peters & Holling 2018).

Fluid intelligence provides an opportunity to investigate the determinants of learning effects in the context of important public health problems.

The most popular test for fluid intelligence is Raven's Standard Progressive Matrices (Raven's SPM) (Kaplan & Saccuzzao 2018; McCallum et al. 2000), and it is suggested that they (including Raven's other progressive matrices tests) are ideal for research into the nature of intelligence (Kline 2000).

The main contribution of this thesis is the design and development of a new automatic item generator, called Corvus, which is used here to investigate the impact of the uniqueness of items on the validity of repeat cognitive testing with Raven's-like tests in an epidemiological context.

Corvus is capable of producing a practically unlimited number of unique Raven's-like test items (Section 4.4, page 95), and the processes of its initial design, development, evaluation and use are detailed in this thesis. It is hoped that all

of these iterative and interacting processes will be ongoing beyond the end of this DPhil.

1.2 Context

There is an increasing need for accurate and valid assessment of cognitive change over time. Current assessments, such as the Mini-Mental State Examination (MMSE) and Alzheimer's Disease Assessment Scale-Cognitive (ADAS-COG) which have regulatory approval for use in clinical practice and trials, provide global measures with pronounced ceiling effects and which are insensitive to early change, and have learning effects on repeat measures (Galasko, Abramson, Corey-Bloom & Thal 1993). These measures are unsuitable for use in population studies where the entire distribution of cognitive performance (not just the lower end) is of interest.

Fluid intelligence is a cognitive domain of specific interest in epidemiological contexts, where it is used in mental deterioration test batteries (Carlesimo et al. 1996; Isella et al. 2003; Rentz et al. 2004), to detect or predict mild cognitive impairment and the dementias (e.g. Cervilla, Prince, Joels, Lovestone & Mann 2004; Harrington et al. 2018), and is linked to a wide range of health issues including hospital admissions, heart disease and physical functioning (Ian, Martha, John, Lawrence & Helen 2014; Singh-Manoux, Ferrie, Lynch & Marmot 2005).

However, learning effects are still present in tests of fluid intelligence, and this presents a significant constraint on research into ageing, such as with the Caerphilly population of men aged 60-75 years who became 'more intelligent' on average over a five year period from baseline to their first retest (Caerphilly Prospective Study, personal communication, 2018).

Conventionally, such learning effects could be addressed by the development of parallel test versions, however this strategy is labour intensive and as a result it is rarely practical to do more than a few retests. An alternative strategy, and the one employed in this thesis, is to design an automatic test item generator that will generate unique test items to a specified difficulty for each instance of the test. This does require computer based testing, however computer based assessment of specific cognitive domains is increasingly used in population studies, such as Gallacher et al. (1999) and *UK Biobank* (2018).

It was initially assumed that intelligence did not change over time – both at an individual level, and at a population level (Raven 2000). This was first corrected at the population level (Flynn 1984; 1987; 1999), and more recently at the individual level, when looking at the effects of outside effects or interventions, such as education (e.g. Skuy et al. 2002), environmental factors (e.g. Dimitriou, Le Cornu Knight & Milton 2015), ageing (e.g. Johnson, Corley, Starr & Deary 2011), and neuropsychological health issues such as dementias (Strauss, Sherman & Otfried Spreen 2006). Nonetheless the assumption still exists in psychology that an individual's underlying intelligence is static (e.g. Villado, Randall & Zimmer 2016), and instead retests are most frequently thought of as a measure of the test's reliability (Joint Committee on Standards for Educational and Psychological Testing 2014), rather than as a measure of change. But that view is not universal, and there is a rapidly growing body of evidence that intelligence is responsive to training (Anastasi 1981), and belief that intelligence is a static construct is harmful (Dweck 2000). Regardless of whether underlying intelligence can change or not, measures such as Raven's SPM are sensitive to participant health factors.

The issue of frequently measuring change in fluid intelligence over time is arguably influenced by the problem of nomic measurement (Chang 2004), which is the problem of developing and assessing the accuracy of a novel tool designed to measure something that did not previously have an accurate means of measurement.

1.3 Raven's-like

Raven's SPM is a test of fluid intelligence, which is the most thoroughly researched non-verbal cognitive measure, with over 1500 published studies involving its use as of 2000 (McCallum et al. 2000). It interacts with various neurological and neuropsychiatric conditions, including the dementias (Harrington et al. 2018; Strauss et al. 2006), and as such versions of it are used to assess mental deterioration (Carlesimo et al. 1996; Isella et al. 2003; Rentz et al. 2004).

Raven's-like tests are similarly tests of fluid intelligence and are inspired or modelled after Raven's Progressive Matrices. They are widely used in population and experimental studies such as the Matrix Pattern Completion test in UK Biobank (which is Cognito's Raven's-like Matrices segment), at least in part because Raven's SPM's licensing fees are impractical for large population studies. Raven's SPM is also lacking in sufficient multiple test forms, which limits its use in epidemiological contexts. Raven's-like tests are sometimes described as Progressive or Figural Matrices test items, though these terms are also used to describe a broader range of tests than the term 'Raven's-like'.

Raven's-like tests are also relatively easy to construct (Raven 2000) and are logical in nature and highly systematic, which makes them well suited to automated construction by computers.

Other sources provide greater detail on what exactly Raven's SPM is and about the cognitive domain it measures (i.e. fluid intelligence), such as Chan (2018), which are not replicated here, as they are not the focus of this thesis.

1.4 Terms and Item Structure

A Raven's-like item is a specific subset of a multiple choice test item. Within the broader domain, the relevant components of a test item are referred to as the item's stem and options (Rodriguez 2005). The stem is a test item's "question", and the options are the choices from which participants are asked to select the right answer. Common variant terms for options in the literature include response options (Wise, Ma & Theaker 2012), alternative answers (Linacre, Chae, Kang & Jeon 2000) and distractors (Guttman & Schlesinger 1967). The term stem is fairly consistent in literature dealing with multiple choice tests in general, however in the particular case of Raven's-like items, the stem is referred to as the matrix, while the options have no consistent and unique terminology specific to the sub-domain (See Table 1.1, pages 19 – 21 for terms and definitions).

Item generators require a taxonomy sufficiently comprehensive to uniquely specify each test item generated; however the variables used by a generator to construct items do not necessarily correspond with psychometrically interesting properties. For example, it is not clear that two test items that are otherwise identical except for one anomalous alternative option, differ by a psychometrically interesting distinction, yet a test generator would need to distinguish between them. Conversely, some potentially psychometrically interesting properties may not be used explicitly by a generator, as they would not necessarily enable the generator to

produce any test items it could not already generate. For example Corvus currently does not take element salience as an input, which is used in Meo, Roberts and Marucci (2007) to refer to how easy the elements are to comprehend, which they manipulate by leveraging familiarity of shape and overlapping the images. However such psychometric properties could still be calculated manually. While in the long term it is important for my Corvus to be able to control items with respect to any interesting psychometric properties directly, such properties have in general been considered less important, unless deemed of interest for a specific study.

J. Raven's own taxonomy of items in Raven's SPM (1936) was dictated by five categories of Rule: "Constant in own row" (Identity), distribution–of–two, distribution–of–three, "quantitative pairwise progression" (a subset of distribution–of–three), and "Figure addition or subtraction" (OR, XOR and possibly ANDNOT) (Carpenter, Just & Shell 1990). Since then, our taxonomy and understanding of Raven's-like items has improved, and has benefited from cross pollination with other subjects; for example it is not really until Ragni, Stahl and Fangmeier (2011) that the Logic Gates (e.g. AND, OR, etc...) were fully identified as such in the literature.

A taxonomy set out by J. Raven for attributes present in the set of options is not readily available. Similarly there are fewer papers that look at answer sets in the context of Raven's-like tests, than those that focus on matrices. This may be because such attributes are harder to spot than those in the Matrices, or perhaps simply because they are a less iconic part of the sub-domain, whose most popular and eponymous test was so named at least partly for their distinctive stems ("Raven's Standard Progressive Matrices").

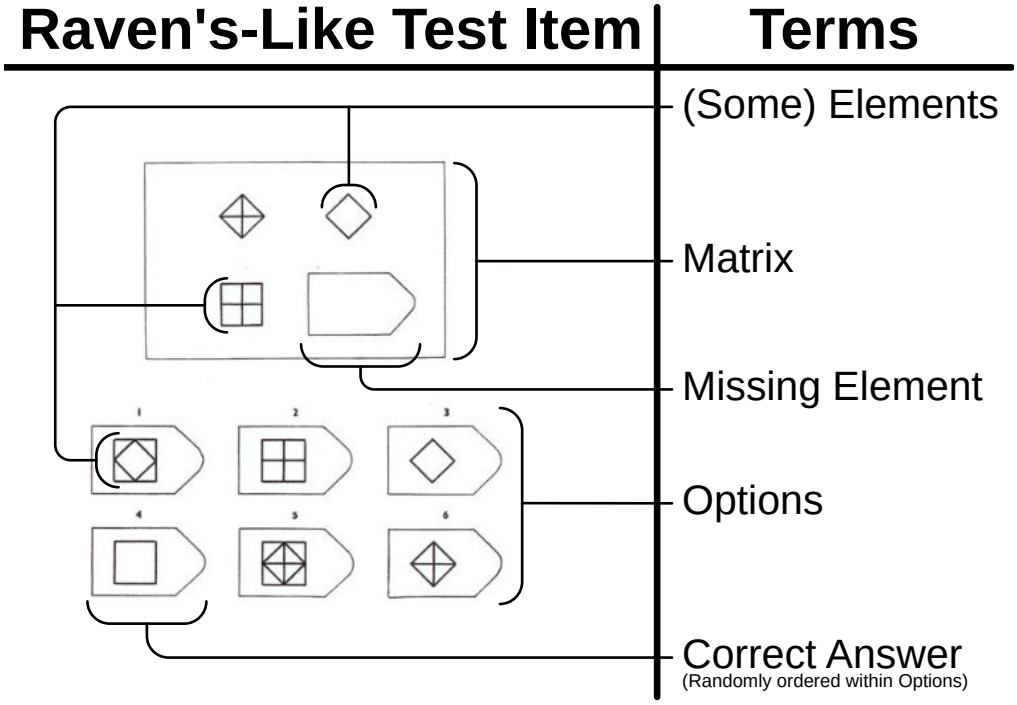


Figure 1.1: An example of a Raven's-like test item, with some terms labeled.

An example Raven's-like test item is presented in Figure 1.1, see Table 1.1 on pages 19 to 21 for more details on the labeled terms. The matrix is defined by Rules or patterns that link the elements of the matrix together. The key, or correct answer, is the missing element from that matrix, which participants must identify from amongst the options presented near the matrix.

A number of terms are used throughout this thesis, that may be subject specific, used in a more specific way than usual, or that are new as they arose out of my analysis in Chapter 3, or from my work in developing Corvus (Chapter 4). These terms are presented and defined in Table 1.1.

Table 1.1: Terms and definitions

Term	Definition
Rule	Though a common term and widely used when referring to Raven's-like test items (and even more broadly), there is some inconsistency in the literature as to what is counted as being part of the Rule or not. In general, the Rule is the underlying pattern that governs the elements in the matrix, and defines the correct answer. However this thesis has opted for a minimal definition of Rule, in that the orientation or layout in which each Rule is applied is not considered to be part of the Rule. This fits with J. Raven's taxonomy of items detailed above. Some examples can be seen in Figure 1.2.
Form	This is a new term, but is used here to refer to the orientation or layout in which the Rule is applied. Common examples of Forms that are used for distribution–of–three or distribution–of–two include horizontally, vertically and the two diagonals. Conversely, Identity has only one Form it can be applied to, which can be seen in Figure 1.3. Other examples of forms can be seen in Figures 4.6 and 4.8 on pages 85 and 86, in Chapter 4, Design of Corvus.
Attribute	Though used here according to the word's normal definition, as part of that, this term is often used specifically to refer to the categories of element attributes acted on by Rules. Examples of Attribute could include shape, size, rotation, colour and number.
Attribute Value	This refers to an element's specific value of an Attribute. For example, a distribution–of–three Rule applied to a horizontal Form, and acting on the shape Attribute, could cause elements in the matrix along the horizontal line, to be variously a triangle, a circle and a square. In this example, triangle is an Attribute value.

Term	Definition
Logic Gate	Logic Gates are a category of functions that take binary inputs and produce a single binary output. They are often described using truth tables, within which the two binary values often use false (0) and true (1). A common example is the function AND, which takes its two inputs and returns false, unless both of its inputs are true, in which case the function returns true. In other words, the function AND is evaluating the statement "x AND y are True" as true or false, where the variables x & y are themselves either true or false, for example the statement "[True] AND [False] are True" is false. In the context of Raven's-like tests the binary values are often represented by the presence or absence of an Attribute, though any Attribute that can handle binary values will do. Three by three matrices handle functions by having two of the elements in a line be considered the input, and the third the function's output. Details on the four Logic Gates presently used in Corvus, including their truth tables, can be found in Table 4.1 on page 78.
Answer Delta	This is a value that measures the number of Attribute Values with which any element differs from the item's correct answer. Some examples are provided in Figure 1.4.
Pattern Delta	This is also a value which measures the number of Attribute Values for Attributes that are not acted on for that item by any Rules other than identity, with which any element differs from all elements in the Matrix. Examples are provided in Section 4.2.3 on page 90.
Anomaly	Any element with a Pattern Delta greater than zero is defined as an anomaly. The closest analogy in White and Zammarelli (1981) would be Oddities, however as they employ a much weaker definition for Oddities, a different term is used here.
Clues & Anti-Clues	A Clue is defined as any attribute for which the answer set as a whole shares a higher frequency of attribute values with the correct answer, than any other individual attribute value, for that attribute. An Anti-Clue is the reverse; for example, when the highest frequency attribute value for a given attribute among the whole answer set differs from the relevant attribute value of the correct answer. Examples and more details are provided in Section 4.2.4 on page 91.

Term	Definition
Option Set	These are the choices presented to participants, from which they are asked to select the correct answer to the test item. They are also referred to as 'distractors' and 'alternative answers' in some literature (e.g. Guttman and Schlesinger 1967). However I preferred options as a term as it generally includes the correct answer, which is not readily apparent with the other terms.
Element	Elements are the constituent parts of a matrix or the choices in an Option Set, each consists of a set of attributes with particular values, such as a triangular shape (where 'triangular' is the value, and 'shape' is the attribute).
Matrix	Matrices are perhaps the most iconic part of Raven's-like tests. The matrix is the grid of elements that form the stem of the test item. The relative position of the elements in the grid is an intrinsic part of the Rules used. In a Raven's like test one element within the matrix is always missing, usually the bottom-right element. The participant is asked to find, understand and apply the Rules governing the relationships between the remaining elements to work out what the missing element should look like. An example of a matrix on its own can be seen in Figure 1.5.
Item	A test item is a specific task test takers are asked to perform. In Raven's-like contexts, this means a matrix together with a set of options to select their solution from.
Correct Answer	This is the option or element in the set of options that best fits the hole left by the missing element in the matrix.

a	a	a
a	a	b
a	b	c

Figure 1.2: Example Rules; each row of this figure is a representation of a different example rule: identity, distribution–of–two and distribution–of–three from top to bottom. Each letter represents a different Attribute value of an element for the Attribute each Rule is assigned to.

a	a	a
a	a	a
a	a	a

Figure 1.3: The only form available to the Identity Rule.

1.5 Corvus

Corvus, a Raven's-like test generator, was designed and built to enable the automated generation of tests with controllable similarity to established tests and novel test items with controllable degrees of uniqueness. The design choices within the test generator were developed through an iterative process that simultaneously informed and was informed by a concurrent formal analysis of established Raven's-like tests, including Raven's SPM itself (Raven 1958) and the Raven's-like test

Figure 1.4: If the correct answer is the far left "black triangle", then the centre element has a Delta of 1, as it differs from the correct answer solely by the colour Attribute and the far right element has a Delta of 2 as it differs both by the colour Attribute and by the shape Attribute.

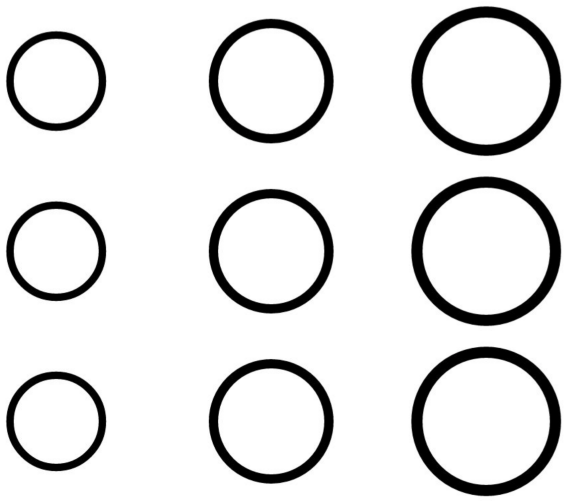


Figure 1.5: An example of a distribution–of–three Rule, with the Horizontal Form applied to the size Attribute. Technically there are also many identity Rules present, but generally they are ignored when defining the matrix if more complex Rules are available. For example the use of circles in this matrix is defined by an identity Rule, laid out in the singular Form available to identity Rules, and applied to the shape Attribute; similarly for colour, rotation and all other attributes.

segment of Cognito (Karen et al. 2014), which is used in UK Biobank (*UK Biobank* 2018). My analysis (see Chapter 3, Focused Literature Review: Comparison of Established Tests) reduced the Raven's-like test items to numerical values determined systematically by their features, in such a way that the original test items could be reconstructed by combining the numerical values with information on the style of graphics used.

Some more complex ideas that were unnecessary to an exact reconstruction of a test item but that either had the potential to be more directly psychometrically interesting or that were mentioned in the literature were also investigated.

There is a substantial methodological difficulty involved in validating a test generator that can generate a practically unlimited number of test items. As a result, and due to the time limitations inherent in a DPhil, a preliminary validation was completed (Chapter 5, Corvus Validation) in which a test generated using

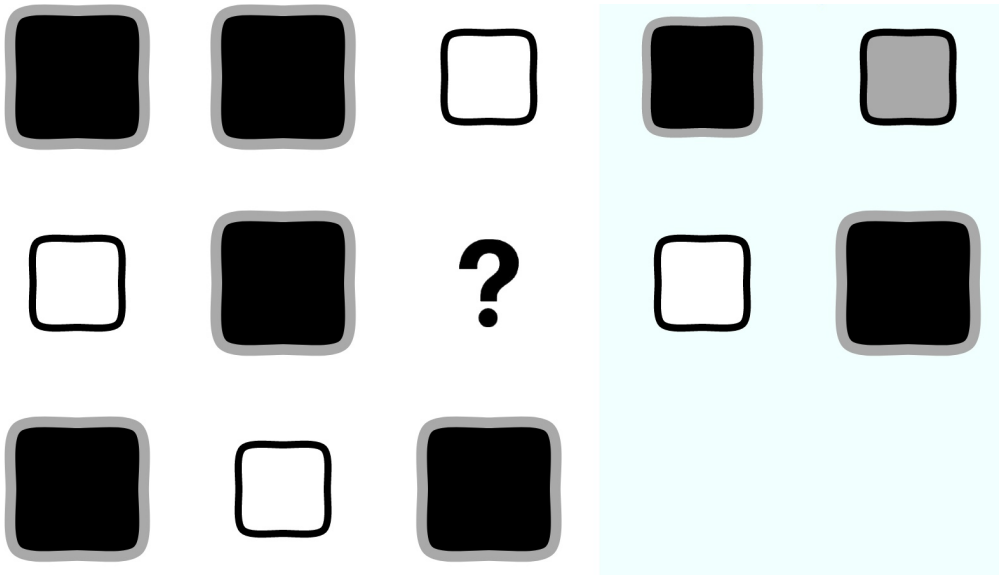


Figure 1.6: An example of a Corvus-generated test item, similar in structure to Figure 1.1, though with the options presented to the right (on the light blue background), rather than below due to computer screens being more often landscape than portrait.

Corvus was compared to other established fluid intelligence tests and was found to correlate well, including with Raven's SPM.

An example of a test item generated by Corvus can be seen in Figure 1.6.

1.6 Adaptive Testing and Item Response Theory

Aside from the potential impact on learning effects, an additional benefit of test generators is their utility in developing test items for use in adaptive testing. This utility is arguably one of the main motivating factors behind the development of most test generators, though Raven's SPM itself continually increased in difficulty as participants progressed through the test. However some modern 'Raven's-like' tests diverge from this, due to adaptive testing.

The model underpinning formal adaptive testing is reliant on Item Response Theory (Wainer et al. 2014), which was designed in response to the limitations and problems in Classical Test Theory (CTT).

According to CTT, test score is an estimate for the underlying trait being measured, albeit that given a certain error that is factored in at the point of interpretation. The primary assumption underpinning CTT is that, for each item, the error in score is uncorrelated with the true score, in other words, the variation in error is equal for all values of the underlying trait. This would mean that every item in the test contributes to understanding the underlying trait in the same way, or to put it another way; there are no misleading items. An implication of this assumption is that the reliability of the test increases with the number of items in the scale (de Ayala 2009). As a result evaluating tests as a whole results in theoretically more reliable results. However, tests developed using CTT have two problems with particular relevance to epidemiological contexts. Firstly, the tests being evaluated as a whole result in a specificity to that test's construction, making even very similar tests (such as different forms of the same test) more difficult to compare. Secondly, as individual error and item–specific differences within the normative samples are not considered part of test score, they cannot then be factored out from the score, which results in a high dependance on the normative sample's characteristics. These two problems make comparisons between tests hard, which causes a third context-specific problem; over time scales are updated and changed, test availability changes, and test appropriateness changes (Streiner, Norman & Cairney 2015). For example, the same test is unlikely to be equally appropriate when the participant is a child, an adult, or if they have mild cognitive impairment. This makes it difficult to compare tests administered at

different times, which makes it difficult determine trajectories over time — and as such CTT is inappropriate for this thesis.

Item Response Theory addressed these problems by evaluating tests on the basis of their individual test items, and more importantly, by building item error into the scoring method – rather than evaluating it afterwards. In other words, instead of assuming that anyone with a certain level of ability would get a test item of a particular level of difficulty correct (assuming zero error when scoring, then applying the error later when interpreting as with CTT), Item Response Theory instead assumes that an individual only has a certain probability of getting each test item correct even if their ability exceeds the difficulty of the item, thus incorporating error into the scoring. This is achieved theoretically by unifying the concepts of test score and the underlying trait being measured into a single measure of performance on the test, which is generally identified by the Greek letter θ (Streiner et al. 2015).

However, Item Response Theory does have its own challenge, in that the amount of work required to produce and evaluate each test is multiplied by the number of items, as each item has to be individually evaluated. At first glance this challenge is insurmountable for Corvus, as it can generate more test items than can be physically tested (see Section 4.4, page 95). However there are two solutions; firstly statistical methods can be used, such as those employed in Geerlings, van der Linden and Glas (2012), which groups large numbers of items together into categories and uses population sampling to make judgements about the items in those groups. The other solution can be seen in Chan (2018), where she uses Corvus to begin investigating the next steps in taking Item Response Theory's concept further and instead of evaluating on the basis of items, evaluating on the basis of test item

attributes in a systematic manner. By approaching the problem from this angle, and narrowing her focus to a specific subset of items and properties, she reduced the number of evaluations needed down to practically achievable levels. Chan was using Corvus to primarily investigate how the figural complexity of elements in items using distribution–of–three applied to a presence-absence attribute value of lines arranged in two annuli impacted test item difficulty. She had mixed results in this goal; finding that figural complexity had both positive and negative correlations with item difficulty depending on various factors. However she did find that her results generally supported Corvus's validity.

While the approach started in Chan (2018) will eventually provide a more detailed validation of Corvus, in the meantime Chapter 5, Corvus Validation provides a pilot validation for the purposes of this thesis.

1.7 Learning Effects

Most studies that specifically target learning effects in fluid intelligence, are primarily interested in the effects of an intervention (e.g. Hausknecht, Halpert, Di Paolo & Moriarty Gerrard 2007a; Skuy et al. 2002). However there are studies that look at repeat cognitive tests without intervention, mostly looking at comparability between baseline and retest scores, and often from education or psychology contexts.

While measuring fluid intelligence at baseline has well-established standards, such as Raven's SPM, the validity of those standards is suspect on repeat testing with intervals of less than between 7 and 13 years (Salthouse, Schroeder & Ferrer 2004).

Test generators produce sets of test items with unique features, yet the underlying structure remains consistent. While handmade tests such as Raven's SPM and Cattell's Culture Fair are generally systematically constructed, they also include some unusual, odd or incorrect items that do not fit the systematic patterns linking their other items.

In either case, although participants may no longer be able to memorise the answers as they would with a conventional test, they may still be able to memorise underlying patterns or methods of solving test items.

This learning of underlying patterns appears to have been part of J. Raven's intent as he designed his test to be a set of overlapping test items, rather than for each item to be wholly novel (Carpenter et al. 1990). This can be seen in the fact that many features of the test items remain constant. Part of the benefit of this is that it makes Raven's-like tests more accessible and relatively language independent; for example, the instructions only need to be explained once. A secondary benefit is that this makes it substantially easier to create item generators.

Raven's SPM's test items could be seen as being too closely related and potentially interfering with the tests' measurement of fluid intelligence, as the items are not wholly novel. However, Raven's SPM has still acquired status as a 'Gold Standard' in measuring fluid intelligence, and its properties have been investigated in thousands of studies (McCallum et al. 2000).

While it is clear that participants having completely memorised answers invalidates fluid intelligence tests, it is less clear how and when that happens, and if memorisation of methods or underlying patterns is also a problem for test validity.

1.8 Main Research Question

For practical purposes, the investigation of a number of research questions are enabled by automatic test generators, including those explored using Corvus by Chan (2018). For my purposes however, Corvus was designed and constructed in a way that also enabled it to answer the specific question:

> *How do test items with varying degrees of similarity or uniqueness interact with the validity of repeatedly administered intelligence tests?*

This question of validity has been highlighted as an area of interest by a number of sources, including Hausknecht, Halpert, Di Paolo and Moriarty Gerrard (2007b); Lievens, Reeve and Heggestad (2007); Scharfen et al. (2018), and some initial steps in investigating it have already been undertaken, such as Villado et al. (2016), and I build on that work here.

Chapter 2

Background Literature Review

2.1 Thesis Focus

This thesis looks to begin addressing the research need to investigate retest validity for Raven's-like tests. As part of this goal, a systematic literature review of retest effects was conducted.

Looking beyond retest effects to the problem of testing Raven's-like retest validity, there is a need for an automatic test generator that can handle varying degrees of emulating traditional Raven's-like tests, especially Raven's SPM. As a result, an investigation of the literature on Raven's-like test design has also been conducted.

2.2 Retest Effects

2.2.1 Mesh terms

The search was conducted primarily on Google Scholar and The University of Oxford SOLO (Search Oxford Libraries Online), with additional checks on other platforms such as Pubmed and Ovid. The majority of references were found via Google Scholar and reference scanning more comprehensive literature reviews. The search terms were formed by using combinations of one or more terms from at least one of the following three categories:

1. Raven's, matrix test, figural matrices, & progressive matrices
2. Learning effect, & practise effect
3. frequent, repetitive, repeat, parallel, & retest

For a flow chart of the literature search and study selection process see Figure 2.1.

2.2.2 Literature Reviews

The initial search identified Scharfen, Peters and Holling (2018) as a relatively recent (as of this review) and highly detailed meta-analysis of a closely related topic — the retest effects of cognitive ability tests [in normal populations] — which screened 34,628 unique records. As such it took a core role in this much smaller literature review. Besides scale, the primary differences between this review and the meta-analysis is that the meta-analysis is interested in a wider range of tests and cognitive measures, but only of healthy participants between 12 and 70 years old, while this review is interested in any adults (18+), whether healthy or not.

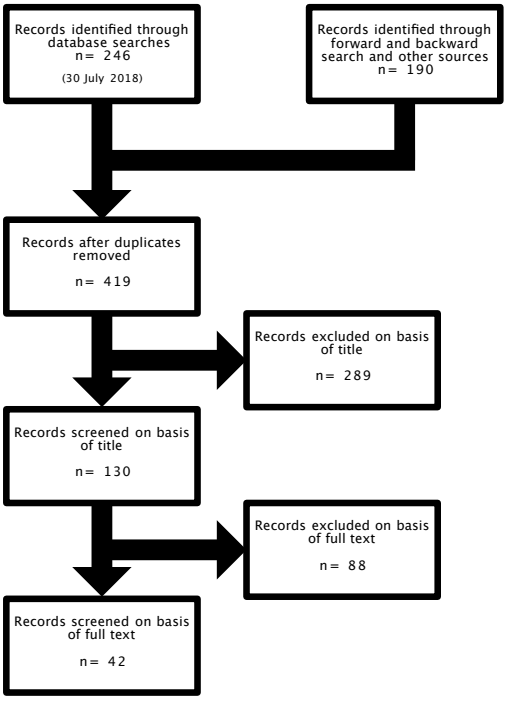


Figure 2.1: Retest effects Systematic Review flow diagram. The initial database searches were on Google Scholar and Solo, with checks in Ovid and Pubmed. However Google Scholar was by far the most useful and identified an up to date and more thorough literature review than was being aimed at here.

Scharfen et al. (2018), identified a number of under-researched areas within the field of retests, including retest effects with a high number of test repetitions (three or more), and how and why retesting changes the validity of a test. Both of those are considered in this thesis, within the specific context of Raven's-like tests.

There are also a number of other older meta-analyses on learning effects, such as Kulik, Kulik and Bangert (1984), Hausknecht et al. (2007b), and Calamia, Markon and Tranel (2012).

Of those, Calamia et al. (2012) was of particular interest for this thesis as it looked at this issue from a clinical perspective, though did not consider the issue of retention of test validity beyond artificial inflation of scores. Nonetheless, a useful finding from that paper was that practise effects for some neuropsychological tests had a smaller impact on score for patients as compared with healthy samples, though unfortunately they were not able to investigate Matrix Reasoning due to a lack of available studies.

McCaffrey, Duff and Westervelt (2000) also provides a literature review from 1970 to 1998 from a clinical perspective, but with the narrow but very practical goal of providing tables detailing the learning effects for a wide range of neuropsychological tests. Although the Flynn Effect limits its application with fluid intelligence for current cohorts, their work is still of potential use for other domains, or around that time period, and for their commentary on the literature in their introduction. The Flynn Effect describes how populations slowly become more intelligent over time (Flynn 1984; 1987; 1999).

By contrast Hausknecht et al. (2007b); Kulik et al. (1984) both took an educational perspective. While Kulik et al. (1984) was primarily interested in methods to help students improve, Hausknecht et al. (2007b) also highlighted the need for research into retest validity.

2.2.3 Primary Literature

A key piece of work on the theory underlying learning effects is by Lievens et al. (2007), who posited that the explanations for retest score changes can be sorted into three main categories:

1. Actual changes in the target construct (e.g. memory tests as a means of improving memory in Roediger & Karpicke 2006)

2. Reduction in contamination; enhancing measurement properties (e.g. reduction in anxiety in Cassady & Gridley 2005)

3. Changes in test-specific, non-general intelligence skills (e.g. test specific practise in Reeve & Lam 2005 and te Nijenhuis, van Vianen & van der Flier 2007)

In the case of the first two, the tests remain potentially useful so long as the results are interpreted correctly. However this is less likely in the case of the third category of explanations.

A large number of papers (e.g. Freund & Holling 2011a; 2011b; Levy & Post 1975; Lievens et al. 2007; Salthouse 2012; Scharfen et al. 2018, etc...) investigate, discuss or otherwise concern themselves with the difference between parallel and identical forms of tests. Although parallel forms are found to have lower practise effects, they do not eliminate the problem of increasing test scores in their entirety (Arendasy & Sommer 2017). Nonetheless there is a common consensus that parallel forms are better when they are available, for example Daniele et al. (2003) states that one of the primary limitations on how frequently they could administer their tests to participants was the lack of sufficient parallel test forms. A common solution is to divide a single test into multiple sub-parts, and use those as parallel forms of each other, for example Estrada, Ferrer, Abad, Román and Colom (2015) separated Raven's Advanced Progressive Matrices (Raven's Advanced Progressive Matrices), amongst others, into two tests by dividing even and odd numbered test items. The other common solution is to wait "a long time"; Bachoud-Lévi et al. (2001) mentions

that the recommended delay for their test battery, including Raven's Coloured Progressive Matrices (Raven's CPM), was six months. Though they increased the interval to a year to try and further limit practise effects. However these solutions only work to a point, and longitudinal studies often look to administer tests a large number of times, for example at the time of writing, the Whitehall Study has twelve phases (*Whitehall II* 2018). Other, generally more recent work, utilises test generators to create parallel forms (e.g. Arendasy & Sommer 2017; Freund & Holling 2011b).

As Scharfen et al. (2018) identifies, studies with more than three retests are relatively rare. Bachoud-Lévi et al. (2001); Bartels, Wegrzyn, Wiedl, Ackermann and Ehrenreich (2010); Salthouse (2012); Staff, Hogan and Whalley (2014) are a few such exceptions, all of which are done either from a health perspective, or with health as a primary motivation or background for their work, a fact that may highlight the relatively unusual nature of epidemiological interest in frequent retesting within the wider scope of research on retests.

Bachoud-Lévi et al. (2001) conducted an interesting study on 22 patients with early Huntington's Disease, retesting them on a range of measures including Raven's CPM, over a period of up to 4 years. Interestingly they often discarded baseline testing when they detected significant retest effects, as the effects were generally largest between baseline and first retest, and discarding baseline helped limit noise from retest effects when investigating disease progression.

Salthouse (2012; 2013; 2015; 2017; etc...) are selected analyses from a larger body of work that draw from the same ongoing study, called the Virginia Cognitive Ageing Project. This study has a 'burst' structure that consists of multiple occasions separated by a number of years, where each occasion includes three

sessions within a single day. In each session participants are administered one of three different forms of the same battery of cognitive tests, including Raven's Advanced Progressive Matrices. Salthouse (2012) found the results of some of their analysis changed depending on if they included or excluded the first session of each occasion. Salthouse (2013) looked at fluid intelligence in particular and found that practise effects for fluid intelligence were influenced by ability and age, while Salthouse (2017) looked at cognition more generally and found that negative ageing effects were relatively small compared with large positive practise effects. Salthouse (2015) combines data from the Virginia Cognitive Ageing Project with data from other studies and observed that the Flynn Effect is a potential problem when looking at longitudinal trajectory of fluid intelligence.

Staff et al. (2014) sought to model the late-life ageing trajectory of Raven's SPM. Their study involved conducting up to five retests of Raven's SPM on 62 to 83 year olds, with the testing intervals ranging from a few months to more than 10 years, with a median of 2 years, for whom they also had childhood Raven's SPM test results, from about age 11. Their study is specifically interesting as an investigation of the competing effects of practise, ageing and the Flynn Effect. They found that participant's Raven's SPM scores decreased on average by half a point per year, with a two point increase on the second attempt, but with an increasingly relevant Flynn Effect in later life.

Other non-health papers do also look at high numbers of retesting, such as Puddey, Mercer, Andrich and Styles (2014), which had access to data from 135000 medical undergraduates' entry exams. While students with three or more retests formed 2.7% of the data set as a whole, they still made a sizeable data set as compared with some other studies. However as participants would only have done

each retest if they failed to get the grades they desired on the previous exam, the paper's relevance to longitudinal cohorts is limited.

The majority of studies, such as Freund and Holling (2011b); Roediger and Karpicke (2006), investigated retest effects, but only for single-retests. There are likely a few causes of this, firstly and most importantly it is cheaper, easier and quicker — if a single retest is sufficient to answer any question, then it is likely this will be the design that will be used. This in turn means a significant amount of the research into retest effects is done from the perspectives of education and psychology, where a higher percentage of their questions can be answered with single retests. This is then self-compounding as it means there are larger communities of researchers investigating the issues in those fields. That said, some work, such as Salthouse (2012), is conducted from a multi-subject perspective.

While such single retest studies were generally of less interest in the present context, some specific studies were still of significant interest, such as Arendasy and Sommer (2017) which investigates the impact of unique test items, and those using Computer Adaptive Testing (CAT) in particular, versus identical test items on single retests. Matton, Vautier and Raufaste (2009) compares the loading of general intelligence of baseline tests with that of single-retests and found that the retests had a lower loading on general intelligence. Freund and Holling (2011b) looked at Raven's-like tests, and administered three tests with the first two on the same day, and the third administered six months later. This was computed in an educational context with 35 classes that were not grouped by the schools on the basis of ability. Freund et al. found that their results were strongly influenced by class memberships; classes that had higher average ability had larger six-month retest effects. This raises concerns that retest effects may be influenced by the

individual's context; for example adults in jobs that push the edges of their ability might have different retest effects to adults who are unemployed. Freund and Holling (2011a) also looked at Raven's-like tests and provided their participants with detailed training before attempting the baseline test, as an attempt to improve group homogeneity. However they found significantly higher retest effects than have been reported in other similar studies, which may have resulted from the pre-test training. Additionally, providing detailed training beforehand may have altered the target construct.

Another important consideration is how long these retest effects endure; however investigating this requires retesting after a significant amount of time, at which point participant retention can become a serious issue. Most investigations into the effect of the length of the interval tend to be either in the region of a year or less, and generally only include one retest (e.g. Arendasy & Sommer 2017; Catron & Thompson 1979; Estrada et al. 2015; Matton et al. 2009), especially when completed without some other means of keeping track of participants, such as their being university students or patients (e.g. Bachoud-Lévi et al. 2001). However Salthouse et al. (2004) was able to source retests with a wide range of intervals, from the following day to 35 years, and detected practise effects with intervals of up to between seven and fourteen years, which is a much larger interval than used in most other studies. Taken as a whole, the data suggest that the long-term impact of practise on test score has a very long tail that decreases broadly asymptotically, tending towards zero with time.

Details of the retest studies can be found in Table 2.1. Some of Salthouse's papers were based on the same continuously recruiting study, and were published at different points in that study's lifetime (2546 participants as of Salthouse 2017),

or used different subsets thereof. Conversely the study detailed in Gallacher et al. (1999) averaged an 85% retention rate each interval of the 2512 participants it recruited at baseline. Also note that for many of the exam based studies, such as Lievens et al. (2007), only applicants who failed every previous attempt at the test are likely to retest.

Table 2.1: Details of studies found on, or with relevance to, repeated cognitive testing

Study	Context	Test type	Size	#Tests	Intervals
Levy and Post (1975)	Health	Test Battery, including Raven's SPM	56	2	4-week
Schaie and Hertzog (1983)	Psychology/ Health	Test Battery, possibly including fluid intelligence	412	3	7-year
Catron and Thompson (1979)	Psychology	WAIS	76	2	1-week, 1-month, 2-months, or 4-months
Giambra et al. (1995)	Health	Test Battery	1721	≤5	Spread over 27.7 years
Gallacher et al. (2009; 1999)	Health	Various, more recently including fluid intelligence	≤2512	3 to 7	5-year
Bors and Vigneau (2001)	Psychology	Raven's Advanced Progressive Matrices	67	3	45-day
Bachoud-Lévi et al. (2001)	Health	Test Battery, including Raven's CPM	22	2 to 4	1-year
Isella et al. (2003)	Health	Test Battery, including Raven's CPM	20	2	1-year
Salthouse et al. (2004)	Employment	Various, sometimes including fluid intelligence	Total: 1110	2	Few days to 35-year
Reeve and Lam (2005)	Psychology	Test Battery, including fluid intelligence	158	3	1-day
Roediger and Karpicke (2006)(a)	Education	Memory	120	2	5-min, 2-day, or 1-week
Roediger and Karpicke (2006)(b)	Education	Memory	180	2	5-min, or 1-week
Lievens et al. (2007)	Education	Course Admission Exam	941	2	2-month
te Nijenhuis et al. (2007)	Psychology	Various fluid intelligence Batteries	176	2	1-day to 3-year
Matton et al. (2009)	Education	Course Admission Exam	752	2	1-year for 90% of the sample
Bartels et al. (2010)	Health	Test Battery, including fluid intelligence	36	5	6-month , or incremental 3-week
Freund and Holling (2011b)	Education/ Psychology	Various, including a Raven's-like	646	2	Same day, then 6-month
Freund and Holling (2011a)	Psychology	Cattell's Culture Fair & another Raven's-like	189	2	1-week or 2-week
Salthouse (2012; 2013; 2017)	Psychology/ Health	Test Battery, including Raven's Advanced Progressive Matrices	≤2546	≤9	Same day or ∼3-year
Puddey et al. (2014)	Education	Course Admission Exam	14739 to 837	2 to 4+	1-year
Staff et al. (2014)	Psychology/ Health	Raven's SPM	751	≤6	∼50-year, then a few months to >10-year
Stafford and Dewar (2014)	Computer Game	Various	854064	≤1000	Various
Hayes et al. (2015)	Psychology	Raven's Advanced Progressive Matrices	35	2	1-week
Estrada et al. (2015)	Psychology	Test Battery, including Raven's Advanced Progressive Matrices	477	4	1-week
Salthouse (2015)(a)	Psychology/ Health	Test Battery	2777	2	7-year
Salthouse (2015)(b)	Psychology/ Health	Test Battery	875	2	5-year
Geyer et al. (2015)	Computer Game	Memory	1890	50+	Various
Arendasy and Sommer (2017)	Psychology	Test Battery, probably including fluid intelligence	960	2	1-month

2.3 Design of Raven's-like Tests

2.3.1 Matrix Design

Carpenter et al. (1990) had access to John Raven's personal notes, and mention J. Raven's taxonomy of his own Rules:

1. 'Constant in own row' (Identity)
2. 'Distribution–of–two'
3. 'Distribution–of–three'
4. 'Quantitative pairwise progression'
5. 'Figure addition or subtraction'

When Raven's SPM was made, there was no established body of work for J. Raven to learn from and this documentation of his underlying concepts used to construct his test items are, from a strict technical perspective, insufficient to construct all of the items in Raven's SPM, yet also contain mathematical redundancies. For example, progression could be considered a subset of distribution–of–three for all test items in Raven's SPM, though this is not generally true for all possible Raven's-like items.

Carpenter et al. (1990) developed a pair of systematic approaches called FAIR-AVEN and BETTERAVEN to solve them and tested those implementations of those systematic approaches in code against human participants. In doing so, they noticed a number of incongruities in Raven's SPM, which I also noticed in my own

analysis (see Chapter 3, Focused Literature Review: Comparison of Established Tests).

Although quantitative progressions are a mathematical subset of distribution–of–three (e.g. 1, 2, 3 is a distribution–of–three and a quantitative progression, while 18, Apple, Circle is also a distribution–of–three; but not obviously a quantitative progression), they have different psychometric properties.

It is also worth noting that J. Raven's description of 'figure addition and subtraction' would be more accurately described using Logic Gates while the final question of Raven's SPM involves true addition and subtraction, which is both mathematically and psychometrically different to other test items that would have probably been placed by him as being under the same category.

The paper, and thus potentially also J. Raven himself, whose comments on his notes did not consider the issue, does not address the impact of answer set design on item difficulty.

The work in Carpenter et al. (1990) provided the basis for a most of the following papers, and they generally fell prey to the same traps. The following papers are looked at chronologically, though starting with one ten years prior to Carpenter et al. (1990).

Mulholland, Pellegrino and Glaser (1980) provides an early systematic investigation on the effect of varying item properties, for geometric analogies (non-Raven's-like, but geometric fluid intelligence items). They found that increasing the number of elements and transformations together resulted in a larger increase in processing time than the sum of parts would have suggested. They also found that increasing the number of transformations correlated with an increasing error rate. However the increases in error rate due to the number of elements were entirely explained by

the increases in the number of transformations (a high number of transformations could require a higher number of elements).

Embretson (1998) developed what was perhaps the first item generator for Raven's-like tests based on the ideas in Carpenter et al. (1990) and validated that the generator could make psychometrically useful tests. She also contributed the ideas of Object Overlay, Object Fusion and Object Distortion as components contributing to matrix definition. Mathematically speaking, her idea of Object Distortion could be considered as an additional Identity Rule applied to the matrix, for example, rotating all elements by a particular angle. Object Overlay and Object Fusion address the issues of interaction between attributes. Object Overlay is when these attributes touch each other, Object Fusion is when the attributes are different facets of the same graphical object, such as the size, shape and rotation of a single object. This work is further built on in Embretson (1999) where she provides more evidence for the validity of Raven's-like test item generators, but for the context of online and adaptive tests.

A second test item generator that incorporated an adaptive engine can be found described in Hornke, Küppers and Etzel (2000), called the Adaptive Matrices Test, and although they identified more rules than most other test designers from this time period, from the perspective of Raven's SPM, there could be considered to be some errors, or non-Raven's-like Rule usage, with some of their provided examples.

For instance the example they provided of their Rule 'Disjunktion' as can be seen in Figure 2.2, which can be solved using the Rule AND, but only if it is applied in the vertical, while applying no Rules at all in the horizontal direction, not even identity. This would break the pattern of Rule usage in Raven's SPM, but does render the example in Figure 2.2 solvable. AND is very close to working in the

horizontal, as only the rightmost diagonal line in the top right element breaks it, which is also part of what has made me wonder if there was a mistake in the example. My attempts to contact the authors have been unsuccessful.

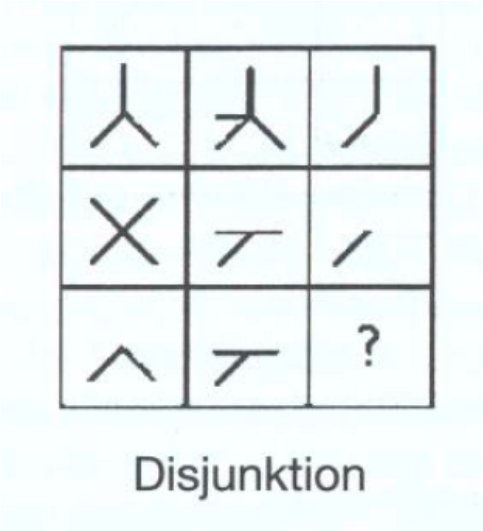


Figure 2.2: Disjunktion as defined in Hornke et al. (2000).

Regardless of their Rule usage, their work provides an early practical example of a small test generator being used with an adaptive engine.

Verguts and Boeck (2002) consider the small number of rules identified by Carpenter et al. (1990) and as a result sought to investigate if and how learning effects apply within the test. Verguts and De Boeck found some learning effects within Raven's SPM, and even stronger ones when participants were provided immediate feedback on correct option selection during the test.

Primi (2002) builds on the work in Mulholland et al. (1980) in investigating the specific sources of complexity in Raven's-like tests. Primi's results show that non-harmonic items (reduced object overlay and object fusion in Embretson's terminology), which are used as a measure of visual complexity, increased test item time disproportionately more than item difficulty. In other words the more

information available that does not contribute towards solving the item, the longer it takes participants to solve it — with relatively small impact on their chance of eventually providing the correct answer.

Arendasy (2005) also produced an item generator, called GeomGen, and added some of the Logic Gates missed by Carpenter et al. though he also seems unaware that they are formally Logic Gates and that he has only identified a small subset of the Logic Gates available. Like Carpenter et al he also misidentifies (not just mislabels) the Logic Gates OR as addition and XOR as subtraction. This may have occurred due to a lack of complete representation somewhere (see Section 4.1.3, page 80). However he has the novel Rule labelled as 'Neighbourhood' in which an element's attributes are defined by a value of its neighbouring elements' attributes (this Neighbourhood rule is the only one not yet implemented in Corvus). GeomGen also deviates slightly from standard Raven's-like tests by providing a 'none of the above' option in the answer sets. This work is further built on in Arendasy and Sommer (2005).

Meo et al. (2007) investigated the potential impact of element salience and identifiability on item difficulty. They found that overlapping and unusual shapes were harder, and suggested that easy-to-identify elements resulted in a lower load on working memory.

As a longer theoretical component of a thesis, Beckmann (2008) was able to include a detailed discussion of developing work on test generators for figural items, though not Raven's-like. Similarly Beckmann was able to discuss validation of her test generator with greater detail than a paper might have, such as how she generated option sets – a topic often neglected. However direct application of her experience with stem generation was limited by the fact that she was working with

a somewhat different test in that regard. Unfortunately her option ('distractor') generation was not optimal, as she assumed that an even spread with a minimal delta across all Attribute values from the correct answer was ideal, a process which creates significant clues as per White and Zammarelli (1981), which is described in more detail on page 49. This occurs as all of the options have all but a small number of Attribute values in common with the correct answer, meaning that the answers can be found without reference to the stem.

Freund, Hofer and Holling (2008) again produced their own item generator, and identified one additional Logic Gate over previous works, i.e. AND, if not by its name. That said, otherwise they used a reduced rule set as compared with their predecessors. More importantly they validated the concept of evaluating item difficulty automatically by working on a subset of the components that go into defining a Raven's-like test item. More complex examples of items made with this generator can be found in Freund and Holling (2011a) and Freund and Holling (2011b), which show four-by-four items using element attributes with no object overlay or fusion (Embretson 1998).

Matzen et al. (2010) conducted a similar analysis of Raven's SPM as that done in chapter 3, Focused Literature Review: Comparison of Established Tests. However theirs was limited to Raven's SPM only, did not look at test attributes as they would have had nothing to compare them to, did not include a detailed analysis of option set design, nor do they address duplicate answer prevention. As a result, items produced by their Sandia generator had to be tweaked by hand in order to prevent poor option sets being used in their study, making it impractical for large scale use. Nonetheless in analysing their own data they did consider option

attributes, if not the attributes of the option set as a whole, and their work is one of few to do as much.

While most other work in the field has been from psychometric or education perspectives, Ragni et al. (2011) approaches the topic from an AI background, and like Matzen et al. (2010) is much more mathematically complete in their definition of matrices. The chief contribution of this work however is the further development of a heuristic for estimating test item difficulty by weighting different kinds of transformations between pairs of elements within the matrix. Ragni and Neubert (2014) takes this work further, includes an analysis of Raven's SPM and the 3rd subtests of the Culture Fair Intelligence Tests, Forms A & B of Scales 2 & 3 (Cattell's Culture Fair), and compares the evaluation of items by computer programs with evaluation by humans.

Arendasy and Sommer (2012) takes a different approach, and looked for the design elements behind gender differences observed in Raven's SPM. Of the three design elements they looked at: Object Salience (the visual prominence of an object), Object Fusion and Object Overlay (see page 43). Of the three only Object Salience was found to have any impact on gender differences and only when a test with a high range of Object Salience was given; a test with constant Object Salience whether high or low was found to have no impact.

Corvus takes a modular approach to object design, and can be tailored with regards to its Object Salience, though the currently produced object modules are primarily low salience.

2.3.2 Option Set Design

Nearly all of the previous papers discussed here do not visibly consider or address the contribution of option set design to item difficulty, as a result the few that do have been particularly highlighted by inclusion in this section.

Rodriguez (2005) presents a meta-analysis of 80 years of research on multiple choice tests and argues that three options are optimal in multiple-choice items. This was due to higher numbers of options increasing testing time disproportionately to the information gained by the item; a result that echoes Mulholland et al. (1980); Primi (2002), but for options. Although it is difficult to argue with his results, it is firstly important to note that the context he is writing for is very distinct from the hyper-specific field of Raven's-like tests, and the following two papers deal more specifically with Raven's-like items. Secondly, it is important for Corvus to be able to make comparisons between itself and other Raven's-like tests more directly. As a result Corvus can generate test items with a wider range in quantity of alternative answers than is suggested as ideal here.

In Guttman and Schlesinger (1967) it is suggested that a good way to design option sets is to look at alternative ways of getting an item wrong, though they do not present a mechanic for doing so with Raven's-like tests, despite spending the majority of their paper addressing them.

However it could be argued that each incorrect attribute is a way for a participant to guess the wrong answer, but presenting the option set as the complete set of all alternative options with an answer delta of 1 (a change from the correct answer of a single attribute value), would present a very clear clue within the answer set to participants.

A potentially better way of applying this particular idea of Guttman and Schelsinger's is to ensure that there is at least one alternative answer that differs from the correct answer per each attribute with an associated non-identity rule. However for harder test items, in practise, this will result in either an inordinate number of available answers or an extremely high average delta from the answer which can also make the test item trivial, as participants would then only need to work out a few of the Rules to eliminate the incorrect options.

Guttman and Schelsinger also assume that anomalies are bad, and fail to notice that anomalies are completely unavoidable in the case of the Identity rule and force test item designers to rely entirely on in-matrix alternative options for matrices with few Rules, which would also make the test item trivial as it would be the only non-in-matrix option.

Conversely White and Zammarelli (1981) incorporated anomalies into their convergence principles, a procedure for finding the answer purely from the option set provided. (White & Zammarelli 1981) showed that their tests, constructed from the answer sets of established fluid intelligence tests, were reliably measuring a cognitive domain but that that cognitive domain may not be fluid intelligence. As a result incorporating anomalies into answer sets should be carefully considered, especially against the alternative of reducing the number of options. However in some cases they are unavoidable, particularly in the case of items solely using identity rules. Meo et al. (2007) is the only paper looked at in this chapter that considers White and Zammarelli (1981).

Arendasy and Sommer (2013) address the issue of alternative options affecting test items, and found evidence that participant methods involving the option set tended to be a fall back method for when solving the Matrix proved too highly

cognitively loaded. They also found evidence that suggested that participant methods involving response elimination lowered test validity, which suggests that answer set design is important. It is worth noting that the methods Arendasy and Sommer address here involving answer sets that are not as developed as those looked at by White and Zammarelli (1981) (they do not reference that paper), and a similar criticism can be levelled at Matzen et al. (2010), which is discussed in the previous section.

As a practical point, both Raven's SPM and Cognito's Raven's-like Matrices segment (Cognito) make extensive use of anomalies, though Cognito to a considerably greater extent than Raven's SPM. As such Corvus needs to be able to handle anomalies in order to emulate such tests.

Table 2.2: Details of studies found on, or with relevance to, the design of Raven's-like tests

Study	Test	Size	Description
Mulholland et al. (1980)	Geometric analogy test items	28	50% female, 18 to 28 years old
Carpenter et al. (1990)(a)	Subsets of Raven's Advanced Progressive Matrices	12	Students
Carpenter et al. (1990)(b)	Subsets of Raven's Advanced Progressive Matrices and Tower of Hanoi split between two sessions	22	Students
Embretson (1998)(a)	Automatically generated Raven's-Like (ART) and Armed Services Vocational Aptitude Battery	728	Air Force recruits, 15% female
Embretson (1998)(b)	Raven's Advanced Progressive Matrices and ART	217	Air Force recruits
Embretson (1999)(a)	ART	818	Young Adults
Embretson (1999)(b)	ART	798	Young Adults
Hornke et al. (2000)	Automatically Generated Raven's-like (AMT)	1236	15 to 77 years old
Verguts and Boeck (2002)(a)	Raven's Advanced Progressive Matrices inspired items, that do not use matrices or multiple-choice	12	Unspecified
Verguts and Boeck (2002)(b)	Raven's-like	16	Persons who 'received course credit'
Primi (2002)	Raven's-like	313	Undergraduates, ∼68% female, 17 to 52 years old
Arendasy (2005)(a);Arendasy and Sommer (2005)(a)	Figural Matrices Generator (GeomGen)	155	51% female, 14 to 57 years old
Arendasy (2005)(b)	GeomGen	897	48.5% female, 13 to 64 years old
Arendasy and Sommer (2005)(b)	GeomGen	451	47.89% female, 13 to 62 years old
Meo et al. (2007)	Subsetst of Raven's SPM and Raven's Advanced Progressive Matrices, plus other Raven's-like items	80	57.75% female, 18 to 34 years old
Beckmann (2008)	Automatically generated figural analogy test items	484	56.2% female, 15 to 20 years old
Freund et al. (2008)	Automatically generated Raven's-like (MatrixDeveloper)	169	66.9% student, 46.2% female, mean age of 27.15 years
Matzen et al. (2010)	Automatically generated Raven's-like (Sandia)	80	Undergraduates, 65% female, 17 to 40 years old
Ragni et al. (2011)	Automatically generated Raven's-like	Unspecified	Unspecified
Arendasy and Sommer (2012)(a)	GeomGen, Raven's Advanced Progressive Matrices Set II, and Numerical Reasoning Test (NID)	620	Students, 50.2% female, 16 to 85 years old
Arendasy and Sommer (2012)(b)	GeomGen, Raven's Advanced Progressive Matrices Set II, and NID	563	Students, 50.3% female, 16 to 85 years old
Arendasy and Sommer (2012)(c)	GeomGen, Raven's Advanced Progressive Matrices Set II, and NID	597	Students, 50.1% female, 17 to 84 years old
Ragni and Neubert (2014)	Raven's Advanced Progressive Matrices, and Raven's-like based on Raven's Advanced Progressive Matrices	17	Students, 64.7% female, mean age of 25.4 years

Chapter 3

Focused Literature Review: Comparison of Established Tests

In this section three Raven's-like tests are quantitatively analysed and compared. The three tests are Raven's SPMs, Test 3 of Cattel's Culture Fair Test (Scales 2 and 3 including both forms A and B, considered as a whole) and Cognito's Matrices test. Wechsler Adult Intelligence Scale, Fourth Edition (WAIS IV)'s "Matrices" segment was also investigated, but due to its inclusion of a significant number of non-Raven's-like test items it was not included in many of the analyses.

Sometimes, in doing the analysis for this review, multiple avenues of describing the same transformation could have been taken. For example, reflection and one hundred and eighty degree rotation of an object that is symmetric in the line perpendicular to the line of reflection both result in identical end positions. In these cases if one of the two methods of manipulation have been used for other purposes in the item, then that method is preferred for describing transformations in situations where multiple transformations could be used. In situations, generally

involving anomalies, where neither of the possible methods are used in any other part of the test item, then the method chosen to describe them is determined by the following order of preference: Number, Size, Colour, Position, Rotation, Mirroring, Translation, Form (Shape), Rules. For example Figure 1.5 on page 23, could technically be considered a change in shape, rather than a change in size — nearly every transformation could be considered a change in shape, which is why it is one of the last in the order of preference. The terms Form and Mirroring are used in place of the words shape and reflection, so that each word starts with a unique letter. This order was decided partly on the basis of Ragni et al. (2011) and partly on the basis of the overall frequency with which that transformation occurs within the tests analysed for this review, while making sure that sub-categories are considered before more general categories that include them.

This chapter and the next make extensive use of the terms defined in Table 1.1, on pages 19 – 21.

Only summary data is presented here due to the necessity of maintaining test security.

3.1 Test details

3.1.1 Raven's Standard Progressive Matrices

Raven's Standard Progressive Matrices (1956 edition, first published in 1936) is historically the most popular fluid intelligence test and is still widely used today, though it has recently been updated with the SPM+. These days the original

test is unfortunately very easy to find online and as such, its security as a test, especially in high stakes environments, is now questionable.

Due to the age of the test, the Flynn Effect is a larger consideration than for most of the other tests considered here.

While Raven's SPM has some problems, it has fewer than most of the others in this review. The most obvious criticism is that there are some test items where arguably weaker cases could be made for officially incorrect options, which is the reverse of the problem with the Cattell's Culture Fair test item. It is theoretically possible that, as has likely happened with Cattell's Culture Fair, were Raven's SPM marked as if these weaker arguments were correct, that those test items would still correlate with intelligence. But this is a criticism that could be applied to all four tests, and is only conveniently solvable via computer-based testing which was not available when Raven's SPM was designed.

3.1.2 Cattell's Culture Fair

Cattell's Culture Fair (Second edition from 1961, first published in 1940) is the second oldest test investigated as part of this review. However the Flynn Effect is less of a concern as to the overall usefulness of the test, as it measured over a much larger spectrum than Raven's SPM was designed to.

Of the tests analysed for this review, Cattell's Culture Fair test presented the most numerous and most significant issues; like Raven's SPMs it is formed from scanned images of hand drawn test items, but both with less precision and at a significantly smaller scale than with Raven's SPMs, which exacerbates the issue arising from the lack of precision. This resulted in several ambiguities, for which

positions needed to be taken for the sake of this review. These ambiguities from graphical imprecision never prevent the correct answer from being identified, but do hinder classification of some of the alternative options. As a result the practical impact of the graphical issues is not critical; mostly they allow participants to more easily discard incorrect answers than they might have otherwise done (for example if two options have no discernible differences, within the bounds of the margin of error of the graphical imprecision, then both can be discarded as answers), resulting in easier test items.

There are test items in Cattell's Culture Fair that do have critical issues, and the publisher has confirmed in email that at least one item is incorrectly marked. However due to the need to protect the test's security, I am unable to go into explicit detail about these test items here.

While individual test items have critical issues, this does not result in an invalid test. Even the incorrectly marked test item will likely correlate with fluid intelligence, because the option marked as correct is a better fit than the other incorrect options, and as a result is a better wrong answer. Cattell's Culture Fair was validated, and these test items apparently performed well enough as is to be included in the final test. This strongly suggests that a non-dichotomous grading system for Raven's-like tests may be worth investigating; i.e. awarding points for how close to the right answer the participant was. It seems possible that such a system could be more robust. This suggestion is reinforced by research in other fields, which has found partial scoring to be superior to 'all or nothing', e.g. Conway et al. (2005).

A third issue is that each scale of the test consists of a Form A and a Form B, which can be done together — however each Form follows the same pattern of

options. In other words, the sequence of answers in Form A and Form B are the same for Test 3. This sub-test structure is unique to Cattell's Culture Fair amongst the tests investigated in this review. Though Raven's SPM has other versions, they were not generally intended to be routinely combined during the same session.

Cattell's Culture Fair does however use a number of forms and attributes not present in any of the other Raven's-like tests investigated. These include a distribution–of–three which defines the attribute acted on by the distribution–of–three Rule in a perpendicular form (e.g. top row is Shape, second row is Colour, third row is Rotation), as well as attributes the other tests do not use, such as forms defined by the intersection of two objects, line dashing and obscured objects. Cattell's Culture Fair also included an entirely novel concept, by hiding elements in the matrix. Generally this works well as in many test items, particularly distributions, there is more information available to the participant than they technically need to solve the item. However, this can remove means of the participant double checking their answer and so may not be a suitable test item property until participants have thoroughly learned the general concepts. It does also increase the lines of symmetry within the matrix, increasing the danger of alternative correct answers.

3.1.3 Cognito

Cognito (2012) is the most recently developed test analysed as part of this review, and the only one of the three looked at in greater detail to have been designed on a computer, or to use colour.

Cognito's main flaw is that it did not take colour blindness into account and a number of its test items fail basic checks for various types of colourblind issues, such as viewing their items through colour filters specifically designed to identify such issues. Otherwise it has few obvious flaws that are unique to it; there are some graphical oddities, such as misaligned shapes, but these are unlikely to have psychological significance.

Cognito's matrices section is the shortest of the three Raven's-like tests analysed here. Cognito is used in UK Biobank under the name 'matrix pattern completion' (*UK Biobank* 2018).

3.1.4 WAIS IV

The Matrices section of WAIS IV (2008) was not looked at in the same depth as the other three tests, as it is not a pure Raven's-like test: it incorporates another style of test mixed in with the matrix items. However it is a popular and established test of fluid intelligence that includes Raven's-like items.

WAIS IV occasionally uses arbitrary attribute values, which can lead to weak test items and in at least one case means that the official answer is not ideal, however as the answer that would fit the stem better is unavailable, this official answer is a superior choice than the alternatives. I have not thoroughly tested WAIS IV for colour blindness issues, but it is quite possible that it would fail such checks, due to its use of colour as an essential attribute in its test items.

3.2 Matrices

3.2.1 Rules

As can be seen in Table 3.1, while Cognito and Raven's SPM are similar in their range and volatility of unique rules, Cattell's Culture Fair and WAIS IV are much more restricted (see Table 3.1), and rely more heavily on other aspects of test item design to provide variation in items and range of item difficulty, such as Forms and Attributes. Additionally, thirteen of Cattell's distribution–of–three items have concealed elements, which is not strictly a Rule, but is a feature unique to Cattell's Culture Fair that seeks to remove some of the unnecessary information present in the matrix by hiding one or two elements, while keeping it still solvable.

Many of the volatility or variability metrics investigated across all three Raven's-like test analyses fit a similar shape to what can be seen in Figure 3.1 when plotted against test item number; relatively flat for the first three quarters, followed by a spike in volatility in the final quarter. It is important to note that Raven's SPM was designed to be administered to children as well as adults (as is Cognito, and Cattell's Culture Fair also). For tests aimed at normal adults, we might have expected higher volatility at the bottom end of the test items, early on in the test, but as Raven's SPM was also administered to children, the test includes more detail at the lower end of the range than a test intended purely for adults might have done. Higher item volatility could also have an effect on item difficulty, as this might reduce any learning effects transferring from one item to the next, and might constitute a second reason to have such a flat curve in the first quarter of Figure 3.1.

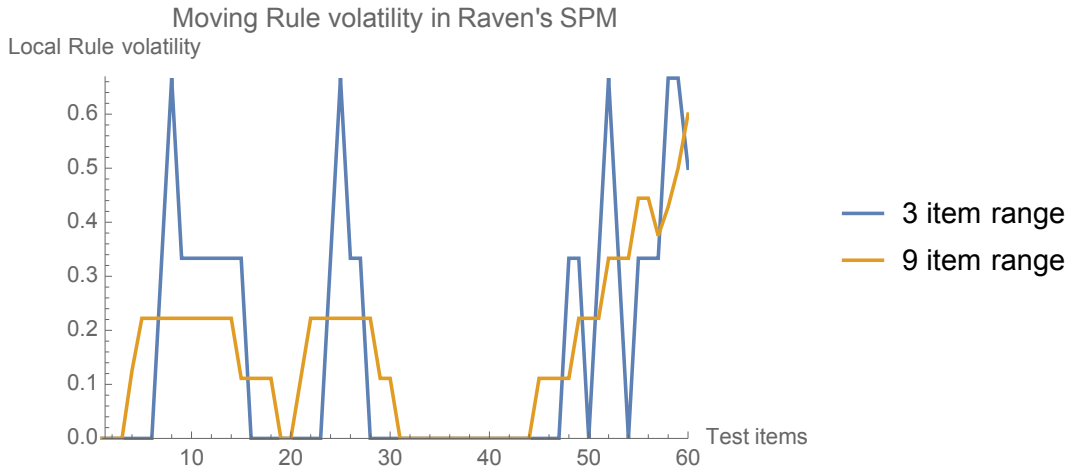


Figure 3.1: This graph shows the mean number of Rule changes within moving windows for Raven's SPM plotted against test item number. The item range is the width or size of each window. The other tests broadly fit a similar pattern. A moving window is a range within which the analysis is limited to, and is repeatedly done as the centre of the window moves along the x-axis, for example there were no changes in rules in a three item window centred on the 30th test item, but 10% of the items in a nine item window, centred in the same place had Rule changes.

That said, Raven's SPM and Cattell's Culture Fair were both created before Item Response Theory became widely used and as a result they may lack a well designed range of items.

A fair number of these analyses also show Cognito as mirroring Raven's SPM very closely in a number of respects, such as in Table 3.1, and I suspect that this was intentionally done by Cognito's creators.

Table 3.1: Use of Rule types

Rules	Raven's	Cognito	Cattell's	WAIS IV
Distribution of 3	40% (24)	46% (7)	32.26% (20)	
Distribution of 2 (2x2)	21.7% (13)	26.7% (4)	62.9% (39)	50% (14)
Identitiy	15% (9)	13.3% (2)	4.84% (3)	3.57% (1)
Distribution of 2 (3x3)	6.7% (4)	6.7% (1)		
OR	5% (3)			
Continuous shapes	3.3% (2)	6.7% (1)		
AND	3.3% (4)			
Combination	1.7% (1)			3.6% (1)
XOR	1.7% (1)	13.3% (2)		
Addition	1.7% (1)			
Distribution of 6		6.7% (1)		
Translation		6.7% (1)		
Layout Implication				3.6% (1)
Not Raven's-like				39.29% (11)

Comparing the frequency of usage of Rule types, note that each test item can use multiple rules – thus the percentages in each column may sum to more than 100%. All of WAIS IV's test items are 2x2, as are some Identity items in the other tests, and all of the distribution–of–two (2x2) items.

3.2.2 Forms

As mentioned in the section on Rules (3.2.1), Cattell's Culture Fair relies on Form (and Attribute) to a greater degree than the other Raven's-like tests and this can be seen in its higher diversity of forms per Rule (Table 3.2).

Size 2x2 matrices have a very low number of possible Rules, and Forms; disproportionally to how frequently the various tests use them, see Table 3.2 (though generally not as disproportionally as Identity and its singular Forms), though WAIS IV has one or two more creative Forms for 2x2 matrices, likely motivated into doing so by the fact that it doesn't have any Raven's-like test items that are not 2x2 (Raven's SPM's use of Combination is in a 3x3 matrix).

3.2.3 Attributes

Unlike the other tests, Cattell's Culture Fair is split into four sections that can be administered separately or combined, thus mechanistic duplicates are perhaps more understandable here than in the other tests, and this can be seen in the high numbers of duplicate test item attributes (Table 3.3).

Table 3.2: The range of Form types for each Raven's-like Rule

Rules	Raven's	Cognito	Cattell's	WAIS IV
Distribution of 3	6 (4)	4 (1.75)	10 (2)	
Distribution of 2 (2x2)	2 (6.5)	1 (4)	3 (13)	2 (7)
Identity	1 (9)	1 (2)	1 (3)	1 (1)
Distribution of 2 (3x3)	4 (1)	1 (1)		
OR	1 (3)			
Continuous shapes	1 (2)	1 (1)		
AND	2 (2)			
Combination	1 (1)			1 (1)
XOR	1 (1)	1 (2)		
Addition	1 (1)			
Distribution of 6		1 (1)		
Translation		1 (1)		
Layout Implication				1 (1)

The number in brackets indicates the average number of test items that combine the same Rule, and same Form – i.e., the only differences between them are due to the attributes or graphical encodings.

Table 3.3: Use of Attribute types

Attributes	Raven's	Cognito	Cattell's	WAIS IV
Size	10% (6)	55.6% (10)	25.8% (16)	3.6% (1)
Shape	36.7% (22)	33.3% (6)	43.5% (27)	32.1% (9)
Rotation/Reflection	10% (6)	16.7% (3)	38.7% (24)	25% (7)
Colour/Shade	11.7% (7)	16.7% (3)	50% (31)	21.4% (6)
Presence/Absence	3.3% (2)	11.1% (2)		7.1% (2)
Number	35% (21)	11.1% (2)	14.5% (9)	3.6% (1)
Height/Width	5% (3)	11.1% (2)		3.6% (1)
Location	36.7% (22)	11.1% (2)	9.7% (6)	10.7% (3)
Rule			4.8% (3)	

This table does not include data from the non-Raven's-like test items in WAIS IV.

3.3 Response Options

3.3.1 Number of Options

As with Rules, Cognito echoes the structure of Raven's SPM with regards to the number of options presented. Cattell's Culture Fair differs slightly, in that it ranges from five to six options, rather than six to eight, while WAIS IV uses five options for all items (Table 3.4). All four tests have more options than are suggested in Rodriguez (2005).

Table 3.4: Number of options presented to participants per test item

Number of Options	Raven's	Cognito	Catell's	WAIS IV
Five			48.4% (30)	100% (28)
Six	40% (24)	33.3% (6)	51.6% (32)	
Seven				
Eight	60% (36)	66.7% (12)		

3.3.2 Delta

The heatmaps in Figure 3.2 are complex to interpret, as the total number of options available changes during all three tests. However much can still be gleaned from them.

A heatmap is a graph with two axes. The intersection of the two axes is coloured according to a legend, usually indicating the quantity or percentage of the measured attribute with the properties indicated by its position on each axis. In the case of the heatmaps on page 65, the property measured is the percentage of test items

in each test, that have both of the properties indicated by that cell's location on each axis. For example in Figure 3.2, for Cognito's Options, out of the eighteen test items there are two test items that have both three Anomalous Options and also two Options that appear in the Matrix.

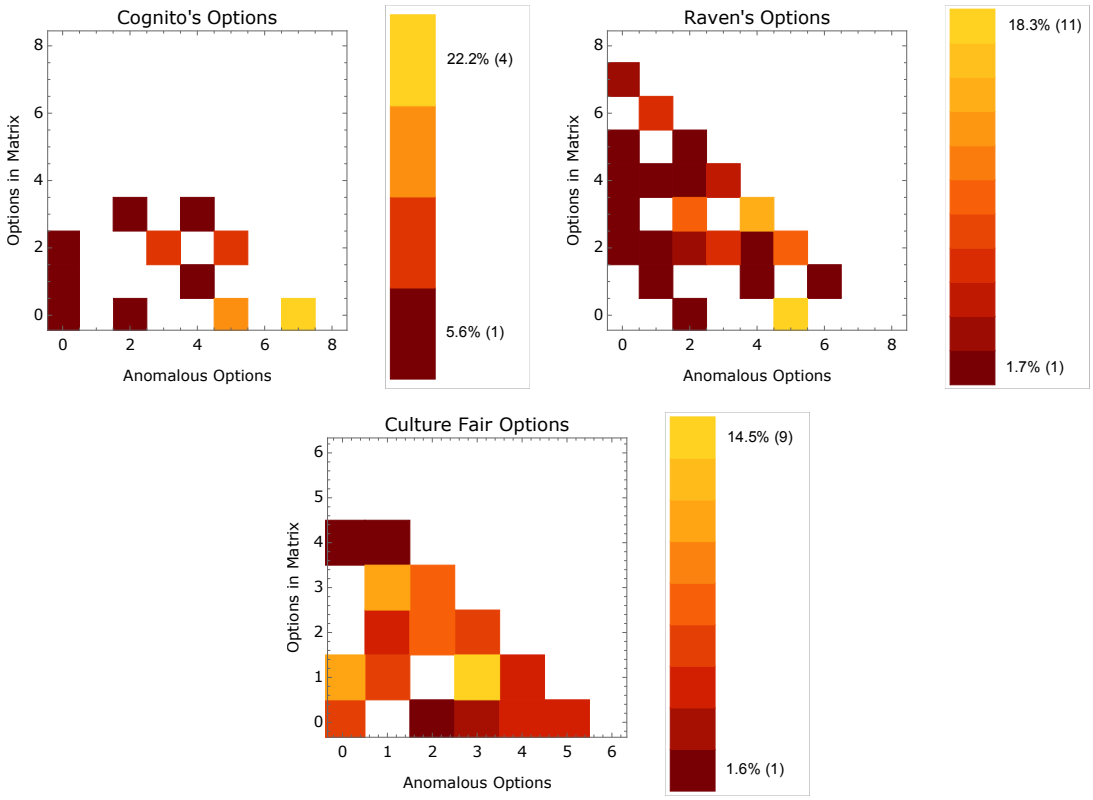


Figure 3.2: Heatmaps plotting anomalous options against options appearing in their matrices for each of the three tests. Yellow colours indicate a higher number of test items, while brown colours indicate fewer, and white space indicate no test items. The exact scales are unique to each graph as there are a different number of test items, and different concentrations of test items in each test, however the percentage and number of items identified by the extremes of colour are identified in each graph's legend.

All three tests have high numbers of test items where all the incorrect options are either anomalous or also appear in the matrix. This can be seen in the hard lines formed where the sum of the anomalous options and options that appear in the matrix equal the total number of options (i.e. the hard edge where the colour

stops in a decreasing diagonal line in the heatmaps for Raven's SPM and Cattell's Culture Fair). These are test items which are guaranteed to be solvable using the method presented in White and Zammarelli (1981) without considering the matrix. This situation is so common in Cognito and Raven's SPM that hard lines can be seen both for the test items with eight options, and also for the test items with six options. The only reason this is harder to spot for Cattell's Culture Fair is that the test varies between five and six options, which as neighbouring integers do not present as clearly separate lines in these graphs.

Secondly, the graphs show interesting information about the relative characteristics of each test. For example, Cognito displays an 'aversion' to presenting options that are duplicates of elements in the matrix, similarly Raven's SPM shows a definite though clearly weaker preference for including anomalous options though it is important to note that it also has four times as many test items as Cognito, thus also having a much wider variation. Cattell's Culture Fair however is relatively balanced in this regard.

Note that identifying if an element contains an anomaly or not is not necessarily as simple as it might seem; particularly with poorly defined or poorly drawn test items. This was not a substantial issue for any of the tests except Cattell's Culture Fair, where a few of the test items were particularly badly made, or even outright wrong. It is arguable that a third category of option should join anomalies and in-matrix options on the basis of some of these Cattell's Culture Fair test items, where there are two options that are functionally identical, however as these were unique to Cattell's Culture Fair, unusual even within Cattell's Culture Fair, and possibly the result of poor drawing rather than intent, these have not been added to the analysis as a separate category. Instead they have been subsumed under

in-matrix options, as the underlying reasoning behind participants being able to devalue them as solutions is arguably similar.

3.3.3 Clues and Anti-Clues

More so than the other tests, Cattell's Culture Fair frequently appeared unaware of Clues and Anti-clues, as it frequently adopted the answer set design of three options using the correct attribute values, two incorrect options using attribute value one and one incorrect option using attribute value two, per attribute. Which results in fairly strong clues being present in the option set.

Raven's SPM and Cattell's Culture Fair are the two Raven's-like tests that predate the first paper on clues in option sets (White & Zammarelli 1981), and are in fact the two tests investigated in the paper. As such, it is reasonable to guess that Cattell was not as aware of this potential issue as he might otherwise have been, and thus did not guard against it. However, as a few of his test items would be unsolvable or dramatically harder without it, it is not clear to me that Cattell's Culture Fair would be better for it.

That said, Clues do not have to be a problem and can instead be viewed as one of many properties available to be tweaked and adjusted in order to produce varying test items to specification. However, it could be that an otherwise very difficult test item, but which possesses strong Clues, is more likely to have its stem ignored in favour of solving it solely via Clues.

3.4 Discussion

Beyond the scope of this thesis, a longer term goal for Corvus is for investigators to be able to specify test characteristics to Corvus – such as those described in this analysis – and have Corvus generate test items, and thus entire tests, matching those characteristics. In addition to providing users with more control, this would also enable Corvus to emulate specific tests without the substantial manual work currently required to do so.

I am not aware of any other work that looks in detail at the characteristics of such tests as a whole, though some such as Matzen et al. (2010) do similar analysis of Raven's SPM alone.

Chapter 4

Design of Corvus

The design and development process of Corvus was heavily informed by an iterative process incorporating the concurrent analysis presented in the previous chapter.

Although the graphical nature of tests has been shown to have impact on test scores (Květon, Jelínek, Voboil & Klimusová 2007), investigating this was not the focus here and has been left for future work; though I did use small pilot studies to ensure that Corvus is colour-blind friendly and has sufficient visual clarity for older participants.

4.1 General Design Choices and Issues

Corvus is a large system consisting of nearly 8000 lines of JavaScript, and was written solely by myself as a component towards the completion of this DPhil.

Corvus is intended for convenient use by participants over the internet. This does mean that due to lag and differences in participant's devices, recorded timings are unlikely to be as precise as in many traditional professional psychometric tests.

However that does not prevent such timings from being useful or interesting, as they will still be precise to a varying degrees depending on the browser and operating system, for example Firefox limits its accuracy to 2ms as a security measure, and some browsers on older versions of Windows, such as Vista and XP, have difficulty counting times that take less than 15ms. Corvus can also be run on the localhost (like an application, though still running in a browser) which could reduce any inaccuracies in timing.

The JavaScript library D3.js is used to draw test items and handle answer selection. Corvus can be placed in a variety of back-ends, and in the studies included in this thesis has been run locally and via the TrueColours self management system. The library JQuery is used minimally with window resizing, and the library seedrandom.js is used to ensure that each test is replicable, given the exact version of Corvus used and the starting point in Unix time in milliseconds (milliseconds elapsed since 00:00:00 UTC Thursday 1st January 1970).

Corvus currently requires browsers that support HTML5, CSS3, SVG and ECMAScript 5, which does not include IE9 and older (i.e., affecting less than <0.98% of users), as well as a few other older browsers that see no use (such as Firefox 3.6 and older), and may not include Opera Mini (<1.18%), which is used almost exclusively in countries where mobile data plans are relatively expensive when compared to local wages. However some of these requirements could be accommodated, and it is likely that Corvus could be altered to support IE9 (<0.25%), though IE8 and older are very likely more effort than they are worth to include NetApplications.com (2018).

Although in general D3.js uses SMIL (which is not supported by any version of IE though work arounds exist), it only does so for animations, which Corvus does

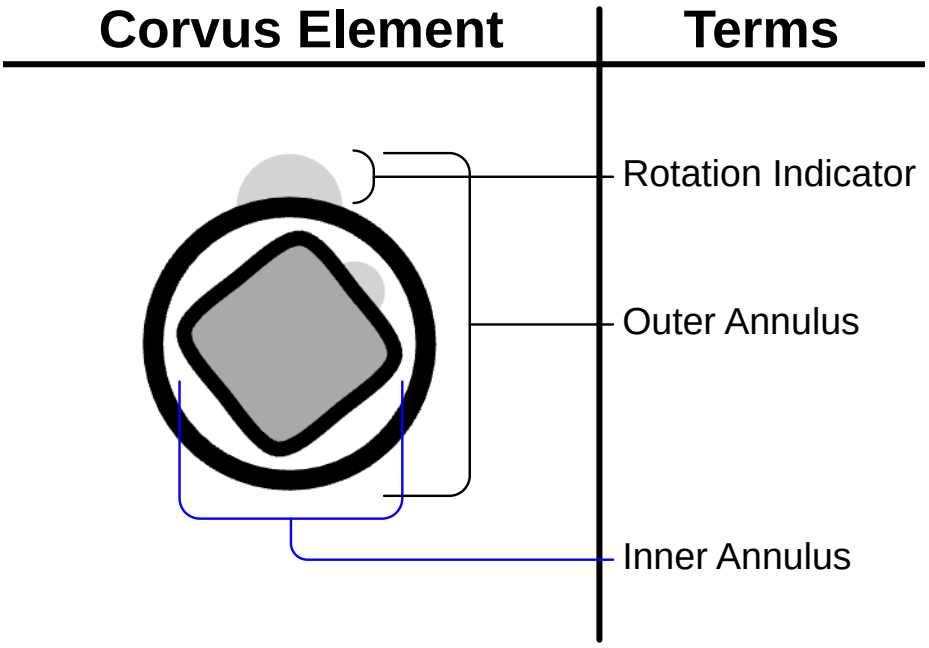


Figure 4.1: An example of an Element as generated by Corvus, with some Corvus specific terms labeled.

not currently use. Since Corvus's requirements, HTML5 etc, are current standards, browser compatibility will improve over time as users update their systems.

4.1.1 Element Structure

Figure 4.1 shows an example of an element in Corvus. Nine similar elements, albeit with one missing, positioned in a three-by-three grid form matrices, such as in the example in Figure 4.2. This example highlights the two-annuli structure of Corvus elements. Separate Rules with their own Forms and acting on different Attributes govern each annulus. When the rotation Attribute is used on an annulus, then a rotation indicator is displayed. This is to prevent issues involving rotational symmetry.

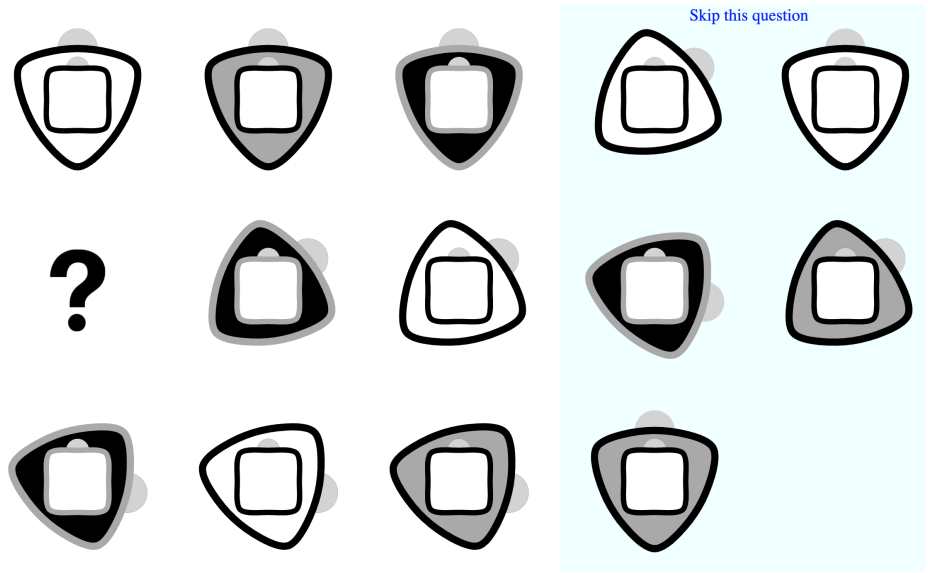


Figure 4.2: An example of a test item generated by Corvus, using the distribution–of–three Rule, in the vertical Form acting on the rotation Attribute, together with a second distribution–of–three in the decreasing diagonal Form acting on the colour Attribute.

4.1.2 Item Structure

Each test item in Corvus is defined by an array and stored in `allPuzzleTypes`. These arrays can be either automatically generated, manually defined (such as the hand-made example using an identity Rule in Code 4.1), or a mixture of the two.

```
1  var allPuzzleTypes =
2          [[[3,3],          // 0. [Grid Size]
3          [0],              // 1. [Graphics]
4          [0],              // 2. [Logic]
5          [0,[0,0,0]],      // 3. [Form]
6          [[0,[1,1,0]],     // 4. [Option Layout]
7          [2,[2,0,0]],
8          [1,[0,2,0]]],
9          [0],              // 5. [Number of elements in centre]
10         [0,0,
```

Figure 4.3: Three example Graphical Encodings, from left to right they are used for; 1) Logic Gates, 2) Distributions & 3) Nesting of graphical encodings for use with Additions and Distributions.

```
11          0,0],              // 6. [Number Layout]
12          [0,0,0],           // 7. [Number Option Layout]
13          0,                 // 8. #Concealed
14          0],                // 9. Type
15          // ... More manually entered test items
```

Code 4.1: A manual test item entry in allPuzzleTypes

Grid Size Raven's SPM progresses from continuous matrix designs, through 2×2 matrices to 3×3 matrices. Corvus currently only handles 3×3 matrices, though most of the code is in place to handle other sizes, including non-square sizes. Though some of Corvus's more complex features would currently have to be disabled with matrices of other sizes.

Graphics Graphic Options is an array of variables that defines the graphical encoding used by the corresponding Logic Option variable. Some examples of available graphical encodings can be seen in Figure 4.3, including one that would have been generated from a Graphic Option (and thus also Logic Option) containing two values, unless one of those Rules was identity.

Logic Like Graphics, Logic is an array of variables, however instead of defining how the numerical values of each element are encoded graphically, Logic Options

define the category of Rule to be used. The full list of Rules available in Corvus at present is presented in Table 4.1 on page 78.

Layout This array defines the Form for basic transformations (magnitude, colour, shape and rotation) and, depending on the number of annuli wanted, either takes the form;

```
[magnitude, [colour, shape, rotation]]
```

or

```
[magnitude, [colour, shape, rotation], [colour, shape, rotation]]
```

Currently for matrices without distribution selected in Logic, Corvus uses identity as the Form relating to basic transformations, and any additional graphical encoding is scaled to fit within the innermost annulus. Generally, combining multiple annuli with an additional graphical encoding and especially with the magnitude Attribute results in poor visual acuity (as the innermost object would be scaled down in size multiple times). Although it is possible to manually define such a test item, Corvus will not automatically generate them.

Option Layout Much like how Layout defines the Form for distributions in the matrix, Option Layout defines the set of options with regards to their properties as they relate to basic transformations. For each Attribute, the Attribute value of the correct answer is defined as 0; thus the correct answer is always either [0,[0,0,0]] or [0,[0,0,0],[0,0,0]], depending on the number of annuli.

The Option Layout is always checked for containing a copy of the correct answer and if it is, an error is thrown. Corvus will not automatically generate an Option

Layout that contains a copy of the correct answer, but such an error could be entered manually.

The correct answer is then added to the array, and the whole array is shuffled, before being used to draw the basic properties of the option set.

Number of elements in centre This variable is used to define the number of binary elements in a graphical encoding that supports them. These elements are essential for Logic Gates, but can also be used for other Rules. The furthest left element in Figure 4.3 on page 73 is an example of such a graphical encoding with this option set to eight binary elements.

Had this option been set to, for example, six binary elements then the basic structure of that encoding would have been preserved (lines connecting dots in a shape around a central dot), but in the outer shape would have been that of a triangle, rather than a square. The structure changes even more when handling odd numbers of binary elements.

Number Layout Number Layout is for numerical transformations, what layout is for basic transformations. It is used to define the Form that the Rule involving numerical Attributes is applied to. Similarly to Layout, in the absence of a defined Rule, identity is used.

Numerical transformations include binary values, such as those acted on by Logic Gates, and the whole numbers as acted on by addition, although in the later case there are usually limits of $-9 \leq x \leq +9$ or similar, for reasons of readability and visual acuity, applied to all elements of the final matrix.

In a 3×3 matrix the four values entered into Number Layout form the four elements that do not intersect with the column or row containing the missing element. The other elements, including the missing element (i.e. the correct answer), are extrapolated from these four.

For the addition Rule, the answer is always equal to the sum of the four values in Number Layout. As the answer must fall within the given bounds for the graphical encoding chosen, this is checked as manually entered values might not meet the criteria, in which case an error is thrown.

Number Option Layout Number Option Layout defines the set of alternative answers with respect to numerical graphical encodings. For binary values, like for Option Layout – and basic transformations, 0 is the Attribute value for each Attribute for the correct answer, however this is not the case for non-binary values.

#Concealed Corvus will attempt to hide elements up to the value of #Concealed, without breaking the matrices complete representation. #Concealed does not work with test items that use any Rule that is not identity, distribution–of–two or distribution–of–three, when it is treated as having the value zero. In a 3x3 matrix it is impossible to maintain complete representation and hide more than four elements, and merely very hard to hide four. In practise, setting this variable to four will often result in only three elements being hidden.

Type This variable is currently unused by Corvus. However it was included with the intent of allowing "odd one out" test items in the future, as opposed to the current "missing element" test items. These are test items where all the matrix elements are present, but one of them does not fit the pattern. Participants are

then asked to find that element; the "odd one out". Such items would not be Raven's-like.

4.1.3 Rules

Rule Definitions

Table 4.1 contains all Rules currently implemented in Corvus and their definitions. Some rules identified by other researchers such as subtraction or progression are considered here as subsets of other rules (for the examples; addition and distribution–of–three respectively).

While it is possible to design answer sets to only include one valid answer so that complete representation is not required, this is not yet implemented in Corvus, as one of the priorities in the development of the engine was ensuring that duplicate answers were impossible, and as a result safeguards against such risks were set in place at multiple levels in the code. It is possible that such answer set design could be a source of providing additional complexity in items in the future.

Unimplemented Rules that could be added to Corvus in the future include the other Logic Gates (NAND, NOR, $p \wedge \neg q, \neg p \vee q, \neg p \wedge q, p \vee \neg q, p, \neg p, q, \neg q$, YES, and NO), Symmetry, Rotational Symmetry, Tessellation, Magic Squares and Karnaugh Maps. Note that some of these would also require or benefit substantially from a wider variety of matrix sizes, for example; Magic Squares are not especially varied when restrained to a 3×3 format, as all Forms for 3×3 Magic Squares are relatively simple transformations of each other.

Table 4.1: Rules in Corvus

Rule	Definition
Identity	x, x, x
Distribution of 2	$x, y, z : x = y,\ y \neq z$
Distribution of 3	$x, y, z : x \neq y \neq z$
Addition	$x, y, f(x, y) : f(x, y) = x + y$ and $x, y \in \mathbb{Z}$
Logic Gates	$x, y, f(x, y) : x, y, f(x, y) \in \{0, 1\}$ The implemented Logic Gates ($f(x, y)$) are defined below.

AND
$x \wedge y$

x	y	$f(x,y)$
0	0	0
0	1	0
1	0	0
1	1	1

OR
$x \vee y$

x	y	$f(x,y)$
0	0	0
0	1	1
1	0	1
1	1	1

XOR
$x \veebar y$

x	y	$f(x,y)$
0	0	0
0	1	1
1	0	1
1	1	0

NXOR
$\neg(x \veebar y)$

x	y	$f(x,y)$
0	0	1
0	1	0
1	0	0
1	1	1

A table of Corvus's Rules for 3×3 matrices. Where x, y, z and $x, y, f(x, y)$ are unordered 3–tuples representing the values of a specific attribute within each element of a sequence (in any given order) defined by the Form the Rule is applied to, such as a row or column.

Matrix Commutativity of Logic Gates

Matrix commutativity of Logic Gates need not be required, however anyone programming a Raven's-like generator needs to be aware of it.

Logic Gates are defined in table 4.1 as $x, y, f(x, y) : x, y, f(x, y) \in \{0, 1\}$. In other words, a Logic Gate is a function that takes two parameters, and outputs a third. Each of those parameters is a binary value (e.g. 0 or 1), in other literature these binary values are often the values true and false, as truth tables make explaining Logic Gate easier. For example, the Logic Gate function AND, gives a result of true only if both of its inputs are true, and false otherwise (i.e. the Logic Gate AND could be read as "are $xANDy$ true?").

A matrix commutative Logic Gate $f(x, y)$ where $x, y, f(x, y) \in \{0, 1\}$ with seed element Attribute values $a, b, c \,\&\, d$ has the layout shown in Figure 4.4.

a	b	$f(a, b)$
c	d	$f(c, d)$
$f(a, c)$	$f(b, d)$	$f(f(a, b), f(c, d)) = f(f(a, c), f(b, d))$

Figure 4.4: Matrix commutativity

Then $\forall\ a, b, c, d \in \{0, 1\} : f(f(a, b), f(c, d)) = f(f(a, c), f(b, d))$. This equation holds true even when the Rules are applied to different Logic Gate Forms.

The matrix commutative Logic Gates are AND, OR, XOR, NXOR, NO and YES.

The other Logic Gates (e.g. NAND and NOR) are not matrix commutative, which means that greater care must be taken when using them, as $\exists\ a, b, c, d \in \{0, 1\} : f(f(a, b), f(c, d)) \neq f(f(a, c), f(b, d))$.

Corvus cannot currently combine two different Logic Gates in different directions while applied to the same Attribute values, however in such cases matrix

commutativity would change to $f(F(a,b),F(c,d)) = F(f(a,c),f(b,d))$ for Logic Gates $f(x,y)$ and $F(x,y)$.

Of the implemented Rules, matrix commutativity is only a concern when applied to Logic Gates, as their matrices are generated by defining a, b, c, d in some particular Form, and extrapolating the rest of the matrix from them. Where as while distribution–of–three, for example, can be thought of similarly (especially when the distribution forms a progression) it is not helpful to generate them like that, and instead they are generated in entirety by applying the Rule to a given Form, as described in more detail in Section 4.1.4 (Distributions).

Addition $f(x,y) = x+y$ where $x, y, f(x,y) \in \mathbb{Z}$, with seed element Attribute values a, b, c, d, is always matrix commutative, as $f(f(a,b),f(c,d)) = f(f(a,c),f(b,d))$ holds true $\forall\ a, b, c, d \in \mathbb{Z}$ as addition is both commutative and associative ($\{f : f$ is commutative and associate$\} \subset \{f : f$ is matrix commutative$\}$).

Complete Representation

Complete representation is required to ensure that test items are solvable if given arbitrary generation of answer sets.

A Rule is completely represented when every aspect of that Rule is fully present within the visible matrix. Logic Gates cannot be completely represented without being applied to at least two Attribute values.

For example, the Logic Gate OR is defined in Table 4.1. For OR to be completely represented in Figure 4.5 matrices, every line of its definition must be visible in its entirety within the matrix. The matrix on the left is not completely represented (as it lacks $OR(0,0)$ and $OR(1,1)$), while the matrix on the right is.

As a result, the matrix on the left in Figure 4.5, could be distribution–of–two, OR, XOR, NAND, YES, or even an unimplemented Rule, such as one based on the line of symmetry in the decreasing diagonal (for which both 0 and 1 would be valid answers).

0	1	1
1	0	1
1	1	?

0, 1	0, 1	0, 1
1, 0	0, 1	1, 1
1, 1	0, 1	?, ?

Figure 4.5: Incomplete representation and complete representation with OR, respectively left to right.

If a test item adheres to the minimum requirements of complete representation, then there is no avenue for participants to check their answers and there are fewer routes to finding the answer. Both of which have the potential to increase test item difficulty.

Rules such as identity, distribution–of–two, distribution–of–three and addition have much shorter definitions, and thus often have much more flexibility with regards to maintaining complete representation.

```
1  function CLGateRNGQu (noEle) {
2    // noEle is the number of binary attributes used in this matrix.
3    // Generate 2x2 matrix, from which the rest of the matrix is
        formed.
4    // This 2x2 matrix need not be the four top left elements,
        although in Raven's they would always be so.
5    // These are instead always the four elements that do not share
        a row or column with the missing answer.
6    var tempQuArray = [];
7    for (var x = 0; x < 4 /*(maxG-1)*(maxG-1)*/; x++) {
8      tempArray1 = [];
```

```
9     tempQuArray.push(tempArray1);
10    for (var z = 0; z < noEle; z++) {
11      tempQuArray[x].push(0);
12    }
13  }
14  // The final array requires at least one of each of the
        following:
15  // 0,0 -> w
16  // 1,0 -> x
17  // 0,1 -> y
18  // 1,1 -> z
19  // It would have at least one of all of these in both the rows
        and the columns.
20  // thus the final array should be considered as four pairs.
21  // First-Second, First-third, Second-last, third-last.
22  // the 'diagonals' need not be considered (first-last & second-
        third)
23  // N.b. Technically this would be sufficient for non-commutative
         \glspl{LG}.
24
25  // This means the following are required (these requirements are
         orthogonal to the array):
26  // [1,1,X,X]
27  // [0,1,X,X]
28  // [1,0,X,X]
29  // [0,0,X,X]
30  // [1,X,1,X]
31  // [0,X,1,X]
32  // [1,X,0,X]
33  // [0,X,0,X]
```

```
34  // [X,1,X,1]
35  // [X,0,X,1]
36  // [X,1,X,0]
37  // [X,0,X,0]
38  // [X,X,1,1]
39  // [X,X,0,1]
40  // [X,X,1,0]
41  // [X,X,0,0]
42  // The orthogonal pairs:
43  var pairs = [[0,1],[0,2],[1,3],[2,3]];
44  // The requirements:
45  var required = [[1,1],[1,0],[0,1],[0,0]];
46  // construction
47  var tempEles = [];
48  var tempValLeft = pairs.length*required.length%noEle;
49  var helpVal = Math.floor(noEle/tempValLeft);
50  for (var x = 0; x < pairs.length*required.length; x++) {
51    if (x > pairs.length*required.length-tempValLeft) {
52      tempEles.push(Math.floor(helpVal));
53      helpVal += noEle/tempValLeft;
54      if (helpVal > noEle) {
55        helpVal -= noEle;
56      }
57    } else {
58      tempEles.push(x%noEle);
59    }
60  }
61  // shuffles the location of the missing element
62  shuffle(tempEles);
63  for (var p = 0; p < pairs.length; p++) {
```

```
64      for (var r = 0; r < required.length; r++) {
65        tempQuArray[pairs[p][0]][tempEles[p+r*pairs.length]] =
              required[r][0];
66        tempQuArray[pairs[p][1]][tempEles[p+r*pairs.length]] =
              required[r][1];
67      }
68    }
69
70    return tempQuArray;
71  }
```

Code 4.2: Corvus's function used to generate Logic Gate Matrices

Concealed Elements Corvus can handle Raven's-like test items with up to 3 or 4 concealed elements (depending on the item, and which elements are concealed) for identity, distribution–of–two and distribution–of–three while maintaining complete representation (Cattell & Cattell 1961).

4.1.4 Forms

Distributions

For distribution–of–two or distribution–of–three in Corvus this means one of four options: Vertical, Horizontal, Decreasing Diagonal or Increasing Diagonal (See Figure 4.6).

These layouts are identified by the direction the Rule is applied in, which is often the perpendicular direction to the line of consistency (a line in which the attribute value in question does not change). For example, the example on the left

in Figure 4.6, is a vertical form, because the Rule distribution–of–three has been applied vertically. This is often unintuitive as the line of constancy is often easier to spot, especially with more complex matrices.

Distribution–of–two uses the same Forms as depicted for distribution–of–three in Figure 4.6, where $a, b, c \in \{1, 2, 3\}$ and $a \neq b \neq c$, but with the modification of $a = b$.

1	1	1
2	2	2
3	3	3

1	2	3
1	2	3
1	2	3

1	3	2
3	2	1
2	1	3

2	1	3
3	2	1
1	3	2

Figure 4.6: Distribution–of–three Forms; note that 1,2 and 3 are arbitrary symbols of specific Attribute values, and do not necessarily form a progression (From left to right: Vertical, Horizontal, Decreasing Diagonal and Increasing Diagonal).

This is not the full list of possible layouts such distributions could take – but by limiting the generator to layouts of a certain kind, we dramatically reduce the possibility of matrices arising with multiple correct answers. Figure 4.7 is an example of such a Form – as drawn it is a Vertical only Form, with no simple Horizontal rule applied at all, and if the bottom right was the missing element, then the test item could have two correct answers due to the near distribution–of–two rule applied in what would also be a problematic version of the Horizontal Form.

Unlike with Logic Gates, the missing element can be placed anywhere in the matrix, though Corvus will seek to prevent lines of symmetry forming, which decreases the probability of the missing element occurring in certain locations, such as the centre, as this can lead to issues with alternative valid answers.

1	3	3
2	2	1
3	1	2

Figure 4.7: Problematic Vertical Form for distribution–of–three.

Logic Gates & Addition

Logic Gates and Addition both take the form $f(x, y) = z$, which requires a more complex range of Forms, though a similar approach and reasoning to limiting their diversity was taken with their Forms, as was with the simpler Rules. These more complex forms are less simply named, and are not depicted verbatim here, though rotating the examples provided in Figure 4.8, can allow the reader to construct all 9 forms.

a	b	
c	d	
		?

a		b
c		d
	?	

a		b
	?	
c		d

Figure 4.8: Logic Gate or Addition Form. The left hand and central Forms can be rotated to construct the other six Forms not depicted. Rotating the right hand Form has no effect as a,b,c & d are arbitrary. As a result there are 9 Logic Gate Forms, which when applied to Logic Gates (but not addition) must have the missing element in the specified location.

Generally given a Logic Gate $f(x, y) = z$ then, with a few exceptions, the function $y = f'(x, z)$ is a multi-valued function and is thus inappropriate for use here, otherwise $f'(a, b) = f(a, b)$ for all of the exceptions (XOR, NXOR, IFq and IF$\neg q$) and the mirror holds true for $x = f''(z, y)$. As a result the example Forms in Figure 4.9 are either not possible for Logic Gates or, when they do work, result in an identical matrix to a Form mentioned in Figure 4.8.

Interestingly if such Forms are used to construct the matrix (even though it results in an identical matrix to one constructed using a Form from Figure 4.8) NXOR looses its matrix commutativity, while XOR retains it.

Addition however is a bijection, and as a result the only restrictions on the placement of the seed Attribute values a, b, c & d in the matrix are that the locations of a, b, c & d are all unique and that at least two further unique locations can be calculated immediately from them.

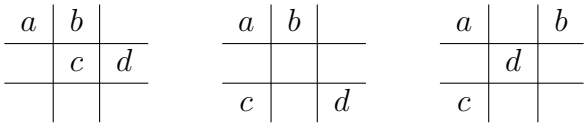


Figure 4.9: Some example Addition only Forms. All of the Forms shown in Figure 4.8 also work, although – as with the examples in this figure – the missing element can be in any empty location.

The only Form used by Raven's SPM for Logic Gates and addition is the furthest left Form in Figure 4.8. Corvus is capable of generating all Forms in Figure 4.8, and it would not be much work to add all of the Forms for addition. However for the purposes of the studies reported in later chapters, Corvus was temporarily limited for Logic Gate and addition to the single Form used in Raven's SPM, which is the far left Form in Figure 4.8.

4.1.5 Attributes

Corvus Attributes include element components such as magnitude (size of element), colour, shape and rotation, integer quantities and binary values (often represented by presence and absence). Early builds of Corvus also included a component that dealt with lines, allowing the generator to play with features such as dashed lines and line thickness. However in early testing it was found that line components caused

visual acuity problems, both with regards to themselves and in compounding issues around spotting size changes (For example the graphical interactions of the attribute values small and thick lines with the size of the overall element – or magnitude, made both tricky to identify). As a result line as an independent Attribute was dropped from the design, and due to the remaining — if lesser — visual acuity issues, relating to the magnitude attribute, Corvus currently only manipulates the magnitude Attribute of the entire element as a single value (including lines), rather than allowing sub-elements to have independent values.

Attributes, and ways to graphically encode them within elements are one of the easiest components of Corvus's repertoire to extend and add to. One particularly interesting Attribute used in Cattell's Culture Fair that could be easily added, is that of the attribute type itself, for example shape, magnitude and rotation could be considered Attribute values of the Attribute, attribute type.

4.2 Option Set Design

Option sets are constructed by choosing the number of options, selecting deltas from the correct answer and from the pattern, and defining the frequency of any alternative option attribute values, so as to control the presence or absence of Clues in the option sets.

4.2.1 Number of Options

Although there is genre agnostic evidence that three options, where the incorrect options are not clearly wrong, is best for multiple-choice items (Rodriguez 2005), and that generally fewer is better, as Corvus generates specifically Raven's-like test

items, a larger number of alternative options than two is usually chosen so as to allow more direct comparisons with other tests in the same sub-genre. Corvus's code however, is flexible in this regard.

Graphical puzzles, such as Raven's-like tests, often have a much higher number of dimensions (For example each Attributes could be considered a dimension) than say puzzles constructed using numbers, which generally only have one dimension. This often leads to much higher overlap between options as a percentage of the total number of dimensions used in the test as options can remain unique and also share multiple attribute values. As a result, Clues and Anti-Clues are a much bigger issue than for many other kinds of multiple-choice tests, and this plays into how the findings in Rodriguez (2005) apply to Raven's-like tests – three is sometimes problematic for Raven's-like tests even if it is optimal in other contexts. Nonetheless Rodriguez's point that minimal numbers of well-chosen options are still more optimal than might have otherwise been considered still remains relevant – especially for simpler test items where fewer dimensions are used.

The domain specific paper Guttman and Schlesinger (1967) is in theory more directly applicable, but it is easy to construct test items for which their guidelines do not work (e.g. Beckmann 2008), or require unpractically large numbers of options. Their paper also lacks the benefit of White and Zammarelli (1981).

4.2.2 Delta from the Answer

Each element, both in the matrix and in the option sets can be measured according to how many Attribute values it possesses that differ from the Attribute values of

the correct answer. This element metric can also be used when analysing matrices, or when comparing options in the option set to elements present in the matrix.

Option sets with lower average deltas relative to the correct answer generally result in harder test items, however Clues present in the option set (See section 4.2.4, Clues and Anti-Clues) become more likely, especially when more than one attribute is being manipulated by Rules and can be unavoidable while maintaining low deltas without providing an impractical number of options.

One potential avenue of future development and a solution to this, is to allow participants to construct their own answer — rather than selecting from a limited selection, though this would increase the risk of valid alternative options.

4.2.3 Delta from the Pattern, In-Matrix Items, and Anomalies

Each element can also be measured according to how many Attribute values it possesses that cannot be generated by the Rules used by the matrix. For example, if all of the Rules used in the matrix are applied to size and shape, then an element that differs in rotation has a pattern delta of 1 – and if it also differed in colour its pattern delta would be 2.

An element could also have a pattern delta of 1 if it varied in shape in a way not accounted for in the Rule. However some caution needs to be applied when considering what can and cannot be accounted for by the Rule, as considering the element under a more general definition of the Rule might reduce its pattern delta to 0.

A small circle in the example used in Figure 4.10, would have a pattern delta of 0. Although all elements that appear in the matrix (referred to as an In-Matrix element) have a pattern delta of 0, it is possible for an element to not be In-Matrix and still have a pattern delta of 0. This usually only occurs when more than two rules are used in a matrix, as it requires that a specific combination of attribute values do not appear in the matrix in the same element. The corollary of this is that for fewer Rules all elements with a pattern delta of 0 are usually In-Matrix elements.

An element with a pattern delta greater than 0 is defined to be an Anomaly. This definition fits with its use in wider literature on option sets in Raven's-like tests, but has the advantage of being more strongly defined (e.g. White and Zammarelli 1981).

1A	2A	3A
1B	2B	3B
1C	2C	?

3A	3B	3C	2C
1A	2A	2B	1C

Figure 4.10: An example of a Matrix using 2 Rules with all of the alternative answers having a pattern Delta of 0. The correct answer is the only available answer that does not also appear in the Matrix.

4.2.4 Clues and Anti-Clues

Existing research has looked at the issue of Clues, that is when hints of the correct answer are resident within the conflux of the set of alternative answers. Unfortunately the proposed solutions for avoiding these situations are impractical when dealing with the set of all possible test items. For example, Guttman and

Schlesinger (1967) effectively proposed providing a handpicked sub-set of answers with a pattern delta of 0. However for test items with low numbers of rules, this would mean that the correct answer is always the only non In-Matrix answer available, and would mean that all available answers are identical, In-Matrix, and correct in the case of a single Identity Rule. They also view Clues within the answer set as being wholly bad, when it is not clear that this is the case. An alternative view would be that Clues and Anti–Clues are simply another tool in Corvus's belt to use in adjusting test items.

The idea of Clues and Anti-Clues was further expanded by White and Zammarelli (1981) who took Raven's and other Raven's-like test items, presented them to participants without their associated matrices, and asked them to solve the questions purely on the basis of the answer sets provided. A Clue is defined as any attribute for which the answer set as a whole shares a higher frequency of attribute values with the correct answer, than any other individual attribute value, for that attribute. An Anti-Clue is the reverse; for example, when the highest frequency attribute value for a given attribute among the whole answer set differs from the relevant attribute value of the correct answer. Clues are more likely to occur when answer sets with low average deltas from the answer are selected (and the reverse is true for high average deltas and Anti–Clues).

Despite this work, the main design goal in designing Corvus is to be able to emulate established tests to controllable degrees. And the work done by White and Zammarelli (1981) predates two of the four tests considered in the previous chapter. As a result, although methods of automatically generating option sets from scratch were investigated, the method implemented was to hard code the deltas used in the option sets of each test item to be emulated, then vary the option set with

regards to the traits discussed in this chapter from that starting point, depending on the degree of emulation desired.

1A	2A	3A
1B	2B	3B
1C	2C	?

3A	3B	3C	3Z
1A	2A	4A	A3

Figure 4.11: An example of a Clue and an Anti-Clue. The correct answer is 3C; amongst the whole answer set the most frequent attribute value for numbers is 3 (a clue) and the most frequent attribute value for letters is A (an Anti–clue). N.b. White and Zammarelli presented a process in their paper that would still have arrived at the right conclusion had 3A not been an available option.

It is worth noting here, that while option sets were a practical necessity for paper based Raven's-like tests, they could be avoided using computer based tests; by allowing the participant to construct their own answers, and in a way that does not rely on drawing skill. However, this idea was not implemented as it would have been a significant departure from traditional Raven's-like tests.

4.3 Unique Answer Checking

A question arises with Figure 4.12: does it use distribution–of–two, in which case the answer is 0, or is it using OR, in which case the answer is 1. If the answer set contains both 0 and 1 as available options, then it could be argued that there are two correct answers.

0	1	1
1	0	1
1	1	?

Figure 4.12: Is this distribution–of–two or the Logic Gate OR?

The problem of identifying a matrix Rule can be considered mathematically similar to the problem of finding a line that passes through a series of data points. Unfortunately an infinite number of curves connect any number of arbitrary points, however many of those curves are considerably more complex than the simplest (An example of this is presented in Figure 4.13). With a matrix those data points can be thought of as the visible lines and columns (i.e. the rows and columns of a matrix that do not intersect with the missing element, or any concealed elements).

```
1  function symmetryChecker() {
2    var symmetric = [];
3    // Vertical line of symmetry.
4    var tempSymmetry = true;
5    for (var y = 0; y < maxG; y++) {
6      if (!equivalency(self.grid[0][y],self.grid[2][y]))
7        tempSymmetry = false;
8    }
9    if (tempSymmetry)
10     symmetric.push("vertical");
11
12   // ... Code checking for other kinds of symmetry
13
14   return symmetric;
15 }
```

Code 4.3: Part of the function used to check for symmetry in the Matrix

Corvus identifies such potential problem items by looking for alternate solutions using its other rules. Additionally, it also does so using symmetries, especially when those lines of symmetry pass through the missing element. When the missing element lies on a line of symmetry, that symmetry could be used to justify arbitrary

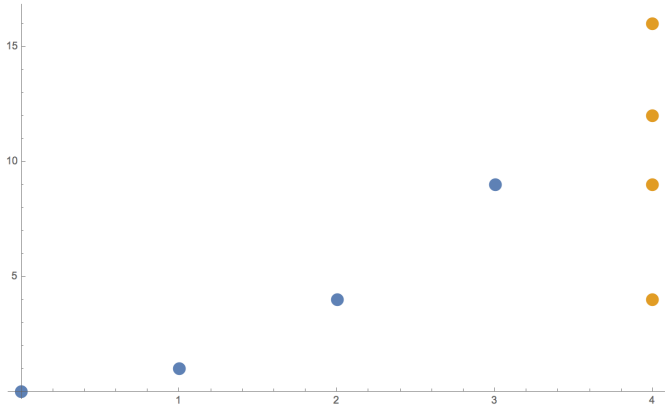


Figure 4.13: The blue dots represent a sequence. Possible next points in that sequence are represented by orange dots.

answers –as any answer would maintain that line of symmetry. As a result, such test items are excluded. Symmetry is also useful as a tool more generally for finding alternative answers, as reflections are by far the simplest transformation not considered by Corvus as a Rule in terms of item construction. This was omitted as none of the tests analysed used it; likely as getting patterns with no arbitrary elements is not trivial for matrices with odd sizes (which cause the lines of symmetry to pass through elements, rather than between them).

4.4 Test Item Capacity

Each feature detailed in this chapter can be combined with each other feature to create a practically unlimited number of unique test items.

Considering the option sets alone, with four attributes (magnitude, with 3 values; colour, with 3 values; shape, with 50 values; and rotation, with 4 values), this gives $3 \times 3 \times 50 \times 4 = 1800$ distinct values for each option. If each of, for example,

4 options takes on those values, there are roughly 4^{1800}, which is approximately 10^{1000} combinations.

Even assuming one minute per item per person, even every human that has ever lived, working collectively from the big bang till now, would not be able to come close to completing this number of test items. So even this low estimate is many orders too large for practical sampling.

Chapter 5

Corvus Validation

5.1 Introduction

Corvus's design was informed by and modelled on the design of established and widely used Raven's-like tests, such as Raven's SPM itself (Raven 1958) and Cognito (Karen et al. 2014), which is used in UK Biobank (*UK Biobank* 2018). Nonetheless it is essential to acquire empirical evidence that the test functions as intended during development. This was done in part to check if the work completed thus far was working, to see if course correction was needed, and additionally as establishing test validity was a necessary initial step before using Corvus later in this thesis in more targeted studies.

A more thorough validation of Corvus is not possible via traditional test validation methods, due to Corvus's capacity to generate a practically unlimited number of test items. Nonetheless this chapter serves as a pilot validation study. However Chan (2018) has recently begun applying new approaches to validation with Corvus, and also found that her results generally supported Corvus's validity

(See section 1.6 for a more detailed discussion of this problem, and further details on Chan's work).

The aim of the pilot study was threefold:

1. To check that the direction of Corvus's development was sound
2. To establish the feasibility of using Corvus (acceptability, functionality, etc.),
3. To provide evidence on the validity of the test generator by the comparison of scores with several widely used tests.

In pursuit of these goals Corvus was compared to a set of established tests in a latin square design. These established tests consisted of a modern, short version of Raven's SPM (Chan 2018; Raven, Prieler & Benesch 2005), Cognito (Karen et al. 2014), and the UK Biobank's Fluid Intelligence test (UK Biobank's test) (*UK Biobank* 2018; Gallacher, personal communication 2018).

5.2 Design

Eligibility criteria were adults (18+ years) without cognitive impairment, who were able to travel to Cardiff University and have normal or corrected to normal vision. Participants were compensated with a £20 Amazon voucher. Using an estimate of association, it was predicted that 27 would be required to detect a correlation of r=0.7 at p=0.01 with 95% power.

To balance order effects, a Latin Square design (Figure 5.1) was used.

Mood was assessed prior to testing using the Mental Health Inventory-5 (MHI-5) questionnaire. MHI-5 was chosen for this as it is a general assessment of mood, it

is not overly long, it can be administered and completed unsupervised, and it is appropriate for use on the internet.

As this study uses a Latin Square design other differences between participants are unlikely to be relevant.

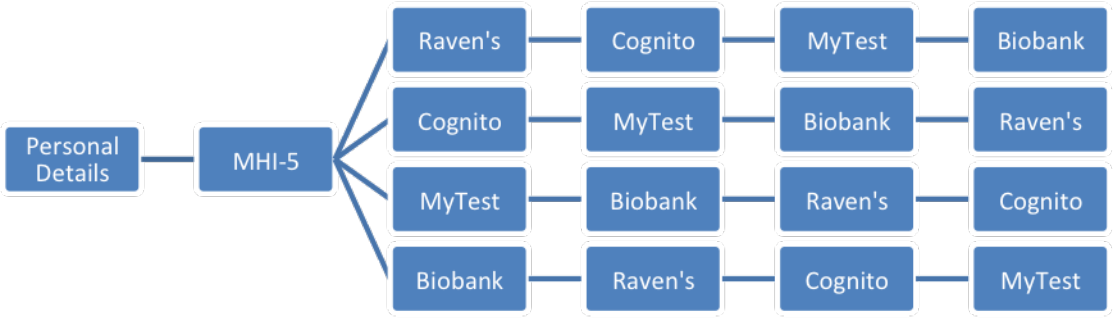


Figure 5.1: Pilot Validation Study Latin Square Design

5.3 Materials and Methods

Corvus was used to generate a set of 16 Raven's-like questions spread across the range of items the test generator was capable of constructing at the time. The set was then placed in an approximate order of difficulty based on the results from a previous informal alpha test study and Ragni et al. (2011). The details of this set of items can be found in Appendix A.2.

Other than art-style and the fact that Corvus's test items have not been formally calibrated, Corvus also differed from Raven's SPM+ and Cognito in two other ways. Firstly the answer set was presented to the right of the Matrix so as to make more efficient use of screen space (Raven's SPM was designed for A4/A5

paper sizes), and secondly mousing over answers or alternative answers graphically inserted them into the matrix, allowing participants to see what the matrix would look like with their answer included.

The short Raven's SPM is an adaptive version of Raven's SPM+, limited to the first 15 questions and built using the open source adaptive engine Concerto (Chan 2018; Raven et al. 2005).

The Cognito matrix test, from the wider Cognito battery, has three practice items followed by 15 test items. Cognito differs from Raven's SPM+ (and Corvus) by the use of colour in its items, and is unlikely to be suitable for use by participants with colour blindness (Figure 5.2). However, none of the participants reported any visual issues with the items.

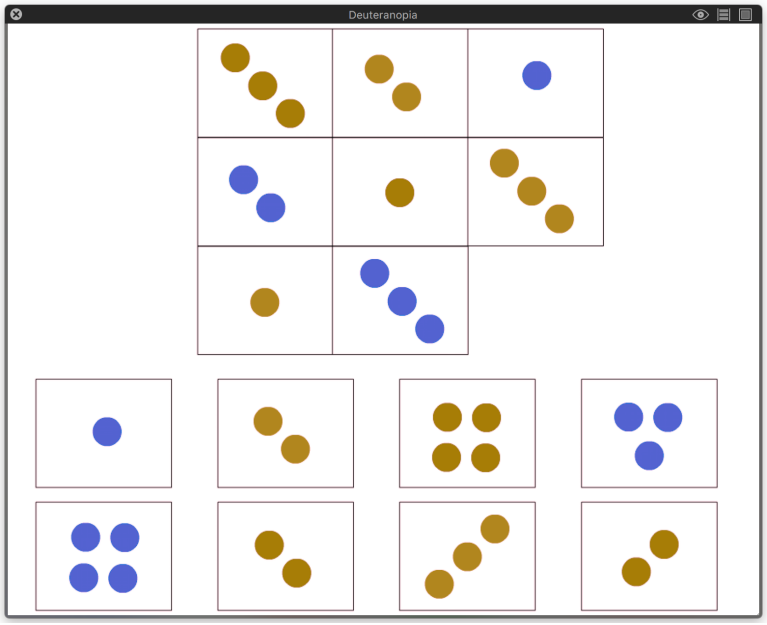


Figure 5.2: Screen shot of a Cognito test item taken through a deuteranopia (Red-Green colour blindness) filter.

UK Biobank's test is designed in a different style to the other tests. It is a two minute timed test in which participants are asked a variety of verbal and numeric

reasoning questions in a similar format to the widely used AH4 test (Heim 1967). The test starts with an untimed initial question, followed by 13 timed questions. Once a question has been started, participants are allowed to finish it even if that takes them beyond the two minute time limit. More details about these tests, questionnaires, the information pack (including processes) and consent form can be found in Appendix A.1, starting on page 157.

The test was completed in a variety of quiet locations that were free from distractions, as was convenient for each participant. This included an office room set aside for the purpose at Cardiff University, various homes and an empty café.

5.4 Results

34 participants were recruited and completed the task, and their descriptive statistics can be found in Figure 5.3.

In a few cases (n = 6), participants accidentally skipped one or two test items. In these cases their scores were marked out of the total that they attempted, rather than the total number of items. This occurred due to the participants double clicking on the correct answer – which was fixed in future studies by switching off the response buttons for 0.5 seconds after loading new test items.

The test generator otherwise performed as intended for all participants, however there was some feedback about the layout of the test items. In particular, the vertical progress bar was not recognised as a progress bar by most participants; also one participant misinterpreted the button allowing participants to 'skip the current test item' as an instruction to do so.

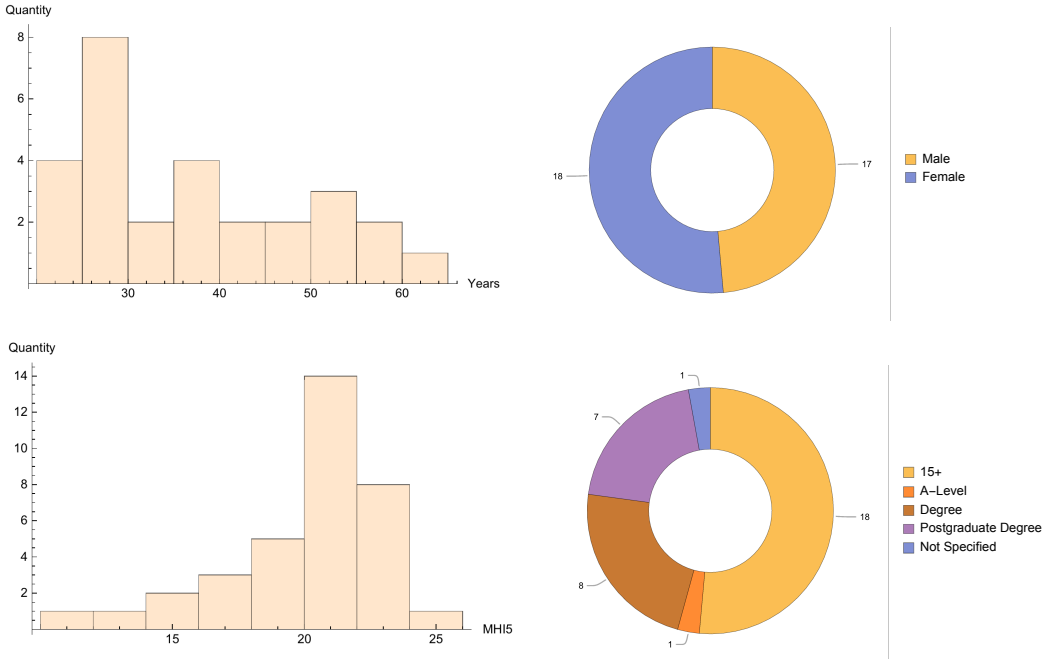


Figure 5.3: Distribution of age, gender, MHI-5 scores and highest education level.

Table 5.1: Correlation matrix

	Corvus_2	Corvus	Raven's	Cognito	Biobank	Age	Education	MHI-5
Corvus_2	1							
Corvus	0.972	1						
Raven's	0.629	0.592	1					
Cognito	0.650	0.662	0.721	1				
Biobank	0.531	0.542	0.593	0.558	1			
Age	-0.187*	-0.304*	-0.009*	-0.079*	0.132*	1		
Education	0.313*	0.297*	0.589	0.548*	0.466	0.090*	1	
MHI-5	0.430*	0.431*	0.382*	0.289*	0.210*	0.119*	0.172*	1

* $p \geq 0.01$
Otherwise $p < 0.01$

It was discovered during the study that the computer program that ran Cognito test failed to save the first participant's data. The cause of this issue could not be diagnosed given the short time available before the next participant was scheduled. As a result a paper version of the test was printed and used for all remaining participants, which had the consequence of removing the test's time limit.

The distributions of the test scores are presented in Figure 5.4. Most of the tests are normally distributed, although UK Biobank's test and Cognito were relatively skewed (Table 5.2), with Cognito also showing a substantial ceiling effect – though this was likely due to its change of format (computer to paper, and speed or timed to power).

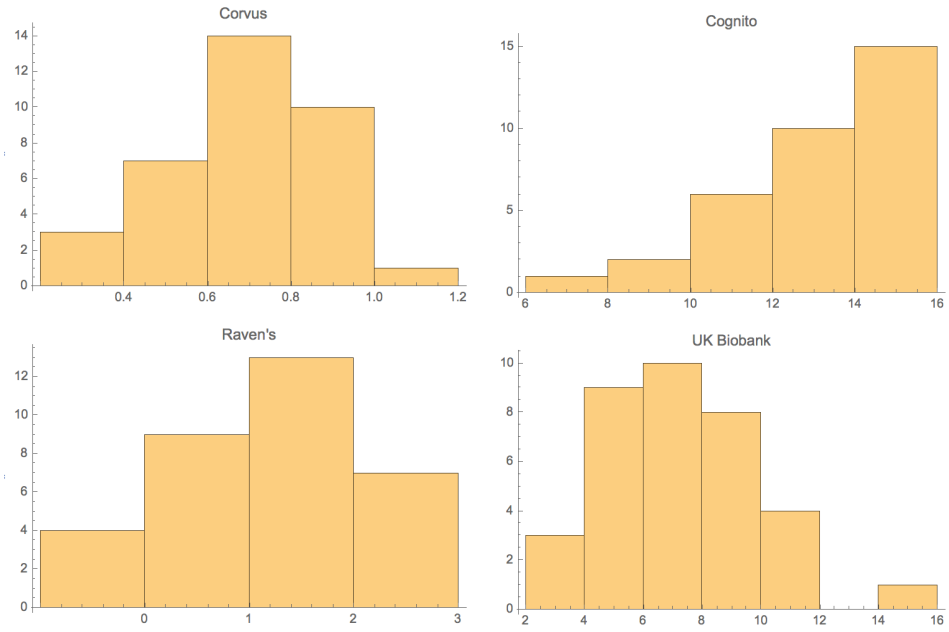


Figure 5.4: Test distributions plotting score (x-axis) against number of participants (y-axis).

The correlations between the new test generator and Raven's SPM+ can be slightly improved by assuming that the population of the study is representative and scaling the score from each question by the percentage of participants who got

Table 5.2: Mean, Standard Deviation, Skewness and Kurtosis

	Mean	SD	Skewness	Kurtosis
Corvus_2	7.3709	2.7505	0.0847	2.4855
Corvus	10.5733	2.7539	-0.3806	2.7629
Raven's	1.1945	0.8664	-0.3482	2.7810
Cognito	12.6471	2.2814	-0.8616	3.2745
Biobank	6.6857	2.5983	0.6473	3.1638

that question correct; in essence awarding 'marks' for each question proportionally to its estimated difficulty. The improvement is small ($\Delta r \approx 0.04$), but is included in Tables 5.2 & 5.1 as well as Figure 5.5, and is labeled as Corvus_2. The reduction in skewness for Corvus_2 seen in Table 5.2 is due to the way this scaling works.

The results presented in Table 5.1 were tested for normality, and all fluid intelligence tests except for Cognito can be modelled by a normal distribution.

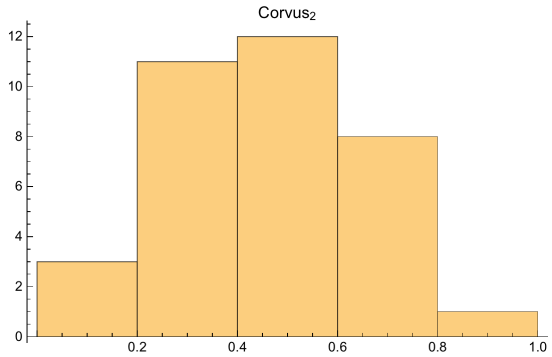


Figure 5.5: Test distributions plotting score (x-axis) against number of participants (y-axis).

The associations between the tests ranged from $r = 0.54$ to 0.72 with all associations being formally statistically significant ($p<0.01$). The strongest association was between Raven's SPM+ and Cognito ($r = 0.72$). Associations between Corvus

and these tests were $r = 0.59$ (Raven's SPM), $r = 0.66$ (Cognito), $r = 0.54$ (UK Biobank's test).

5.5 Discussion

Broadly the primary outcomes for the study were:

- Verification that Corvus's development is heading in the right direction.
- The experience gained through the process of running the test.
- Quality assurance testing for the generator.
- Computing the correlation (r) between test scores.

This study demonstrated that the item generator could generate test items that are acceptable to participants (all participants completed the test).

Evidence on the validity of the generated items is encouraging but raises several issues. Associations with the item generator were lower than might have been expected. Typically test-retest correlations of $r = 0.8$ are reported for Raven's-like tests (Goldstein & Hersen 2000; Raven, Court & Raven Jr. 1983).

There are a number of possible causes for the lower than expected associations. Firstly the new test's unique mouseover feature marks a significant step away from traditional Raven's-like test item design and complicates direct comparisons. Secondly, while the Latin-Square design increases the accuracy of the data, it also increases the variance of the attribute being measured. Due to the small sample size of this validation study, I was unable to compensate for this. I would have required approximately four times as many participants to do so properly, while maintaining a Latin–Square design with four branches.

Additionally, the new test had not been crafted in the normal way, as Raven's SPM+ and Cognito will have been. This process of manually crafting a new test involves creating considerably more test items than needed and then pruning out the less useful ones until only the test items that appear in the final test remain.

However, the low correlation between Cognito and Raven's SPM+ (0.721 is still lower than expected) is probably hindered by the fact that Cognito's was designed to include a time limit per item, the removal of which potentially resulted in its ceiling effect, but also marks a significant departure from Raven's SPM's design.

The fact that there were a number of design issues in the study is a useful point of learning for future studies.

Few of the test results had statistically significant correlations with age, education or mood. However, all of those associations were in the directions expected, though varying in strength. It may be of note that Corvus's association with age and mood appears slightly stronger than Raven's SPM or Cognito, while its association with education appears slightly weaker. Nevertheless, it was concluded that there is sufficient evidence that the item generator does assess fluid intelligence to warrant further development.

Although not statistically significant, the correlations between test scores and mood were stronger than expected. As a result semipartial correlations were calculated correcting for mood (and showed a mean decrease in correlation of 5%), and the correlations continued to support the conclusion that the new item generator assesses fluid intelligence. These results highlight the importance of mood when testing fluid intelligence.

The next two studies build on the evidence for Corvus's validity via the positive correlations between test scores and education, the specifics of which are detailed

in the results chapter of each study. Nonetheless the methods developed in Chan (2018) are important for the practical and detailed validation of any test generator, as well as for understanding Raven's-like tests in general, and her work contributes to the process of gathering evidence for Corvus's validity.

Chapter 6

Learning Effects Study

6.1 Introduction

Learning effects in cognitive tests can last up to 14 years (Salthouse et al. 2004), but there is often cause to want a higher level of detail when looking at change within participants more frequently than once per 14 years. It is not clear that repeat cognitive tests completed within that time frame are still valid measures, particularly as psychometric validity testing is not generally done for repeat testing contexts.

Perhaps the most relevant paper to this chapter is Villado et al. (2016), who investigated retests of Raven's Advanced Progressive Matrices with a six week interval between the first and second retest, and no interval between the second and third. To do so, they split each of the two forms of the test into two 18-item tests. Each participant was then assigned to one of the four 18-item tests to complete in their first assessment, and then proceeded to alternate forms or identical forms of the test according to a latin square design – with each assessment

given 12 minutes. Along side Raven's they simultaneously assessed participants using similarly processed versions of Wonderlic PT, and also acquired self reported estimates of general mental ability.

They found no significant differences in mean score between test forms, however to compare scores across assessments they found it necessary to standardise their results. They found significantly higher retest effects on score for Raven's than for self reported general mental ability, but significantly lower than the retest effects for Wonderlic PT. They did not investigate time taken, as a fixed amount of time was provided per test, nor did they investigate time taken per test item.

The primary aim of this study is to directly address my main research question:

> *How do test items with varying degrees of similarity or uniqueness interact with the validity of repeatedly administered intelligence tests?*

To investigate this question participants were split into three groups, and asked to complete five sessions of testing using tests generated by Corvus, or an online version of Raven's SPM, with increasing intervals. Each of the three branches had a varying degree of similarity or uniqueness. Participants' responses and time taken were recorded for each item. The most interesting results related to time taken per item, which had changed dramatically by the fifth session.

6.2 Design

After being fully informed regarding the study, and signing consent forms, participants were given the option to provide descriptive data. The descriptive data included gender, age, level of education and mental health. Education was recorded

on a scale of 1 to 5, ranging from the participant's highest level of qualification being GCSEs (1) to having completed a PhD (4). A response recorded as 5 indicated a preference not to say, although no participants chose to do so in this study. Mental Health was measured using MHI-5.

Participants were then assigned randomly to one of the three branches of the study. Corvus_2 was not available initially, so as participants signed up to the study they were alternated between the two available branches. When Corvus_2 was ready to be utilised all participants were assigned to it, barring the final three participants who were assigned to the Raven's SPM branch, due to a lower retention rate in the Raven's SPM branch as compared to the Corvus_1 branch.

The three branches were;

1. Raven's SPM
2. Corvus_1
3. Corvus_2

Each of the three branches consisted of five sessions completing a sixty item Raven's-like test according to the branch of the study they had been assigned to. These five sessions were spread out over two weeks with incrementing intervals (Days 1, 2, 4, 8, and 14), starting on the day they completed the first session. Participants were reminded by email on the day, and the following day if they had not completed that session already. No further reminders were sent.

Corvus_1 and Corvus_2 were both generated using Corvus, while the Raven's SPM branch was an online version of Raven's SPM. Corvus_1 closely emulated Raven's SPM, Corvus_2 also emulated Raven's SPM, but relatively loosely (these tests are described in more detail in Section 6.3, page 111).

6.3 Materials and Methods

The online version of Raven's SPMs was made to be as close as was practical to the traditional written form. As such, it maintained features such as the portrait layout, hand-drawn items and layout of options.

Corvus_1 and Corvus_2 both used tests generated by Corvus, which did not differ from any other Corvus generated tests in terms of layout or art-style. All of the tests used in this study were generated on the fly for each participant and each session. What distinguishes between the tests generated by Corvus for the Corvus_1 and Corvus_2 branches of the study however, are the parameters within which the tests were generated. Participants in Corvus_1 were given tests that very closely emulated the test items used in Raven's SPM in terms of Forms, Rules, and option sets. Specifically the Form was allowed only to rotate by zero or ninety degrees for distributions (e.g. horizontal could remain the same or become vertical), while the Logic Gate and addition forms were limited to having the missing item in the bottom right (as per Raven's SPM). The delta pattern for the options was fixed in general according to Raven's SPM, but not specifically to each Rule where there were multiple Rules, and the order of the options was shuffled randomly. Similarly the Attributes were allowed to vary freely. The tests generated for Corvus_2 only maintained the Rules used by Raven's SPM.

The option sets for both Corvus_1 and Corvus_2 were both manually adjusted to emulate Raven's SPM with regards to deltas, anomalies and Clues.

At the time of running this study, there were two Raven's SPM items that deviated from Raven's normal item categories and rule sets in ways that Corvus was not capable of generating (C2 and E7). For these items Corvus instead generated

items that were unusual in a way that it could generate; in both cases this was achieved by using the concealed element mechanism, as is used in Cattell's Culture Fair.

6.4 Results

53 participants were recruited to take part in this study, primarily via social media. The participants were 18 – 69 years old, had normal or corrected to normal vision, access to the internet via a computer with a mouse or trackpad, and had English as their first language. Additional details can be found in Table 6.1.

The correlations between education and test score were $r = 0.67$ for Raven's SPM, $r = 0.51$ for Corvus_1, and $r = 0.78$ for Corvus_2.

Table 6.1: Study participant descriptive statistics

	Gender		Age		Education		MHI-5	
Study Branch	M	F	Mean	SD	Mean	SD	Mean	SD
Raven's SPM	5	8	35.85	15.01	1.92	0.64	60.92	14.89
Corvus_1	3	10	40.85	15.24	2.23	0.93	75.69	13.01
Corvus_2	10	17	40.04	13.22	2.04	1.06	76.59	10.50

Rank order correlation (Figure 6.1) is well maintained for Corvus_2 and Raven's SPM, but distinctly less so for Corvus_1.

There were two large outliers regarding the length of time one Corvus_1 participant took for the 2nd and 7th test item in their first session. The outliers were of a magnitude and uniqueness within the whole data set that it was felt that the participant was likely distracted from the test during those items. For the sake of analysis the time data for those two items was adjusted so that the participant

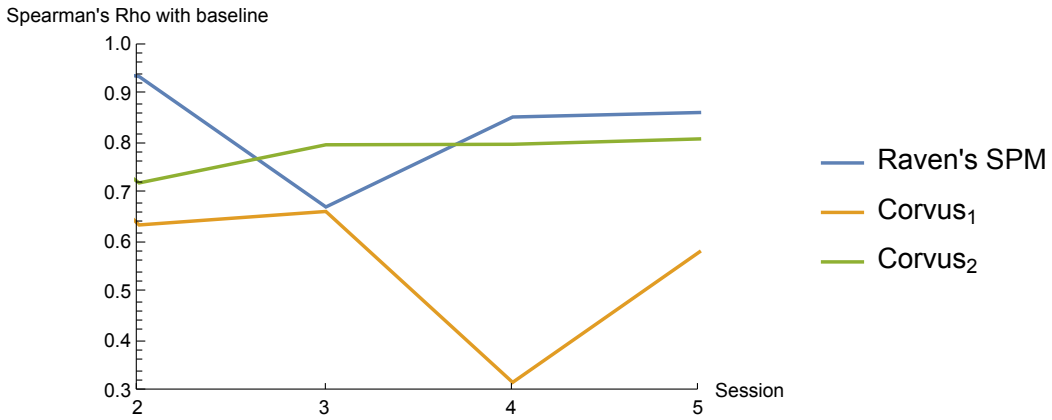


Figure 6.1: Spearman's ρ between session 1 (baseline) scores and scores from subsequent sessions.

took the average of the time they took for the test items either side of the outlying items. The subsequent test item time stamps were also then adjusted for analysis. A more cautious approach would have been to discard the participant, but the study is small and it seemed a reasonable adjustment.

The study showed very little change in terms of score for both between participants, and within participants, regardless of study branch, with mean standard deviation between branches of 4.65 correct answers for 60 item tests (see Figure 6.2). The variation drops even further when limiting the comparison across sessions, as on average there is a slight improvement of about 0.7 correct answers per session within participants.

Test data relating to how long participants took to answer questions was far more varied and interesting (see Figure 6.3).

The most dramatic variation between branches can be seen in Figure 6.3, where in Raven's SPM time taken per test time tends towards a linear relationship with test items on repeat measures, meaning that difficult items take participants no longer than easy items.

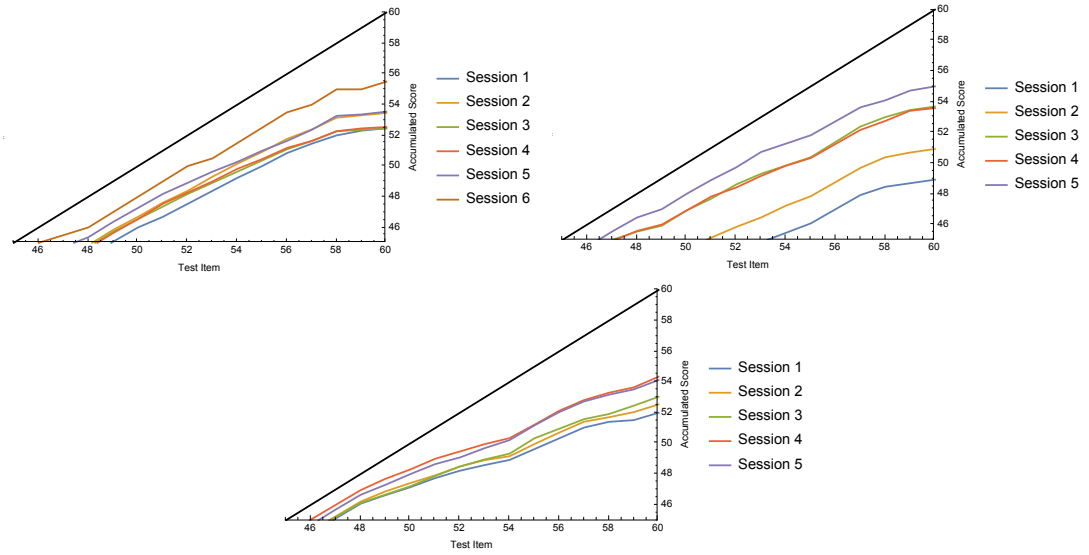


Figure 6.2: Mean accumulated score per session for Raven's SPM (left), Corvus$_1$ (right) and Corvus$_2$ (bottom). The thick, straight, black lines delineate the maximum score per test item (what a participant's graph would look like if they got every item correct).

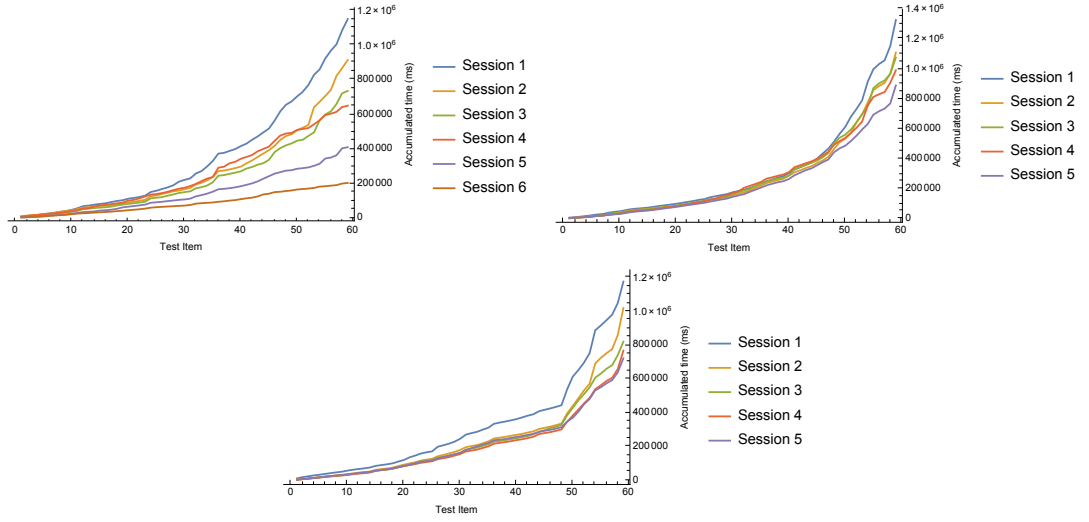


Figure 6.3: Mean accumulated time per session for Raven's SPM (left), Corvus$_1$ (right) and Corvus$_2$ (bottom).

Table 6.2 shows the means of the total time taken for the test, the time taken for the first test item, and for the last test item per test, and per session. By the time of the sixth session the entire sixty item Raven's SPM test is completed on average in ∼202000ms, or 3.37 minutes. In comparison the first Raven's SPM test took 19.2 minutes on average, and the final individual test item in the first session of Corvus_1 took participants an average of nearly 3 minutes on its own!

Table 6.2: Summary data for time taken per session

Test	Measure	1st	2nd	3rd	4th	5th	6th
Raven's	Total	19.20m	15.25m	12.26m	10.83m	6.83m	3.37m
	First	3.06s	2.11s	2.54s	6.23s	1.48s	1.28s
	Last	65.13s	47.17s	14.02s	7.67s	6.55s	3.52s
Corvus_1	Total	22.11m	18.49m	17.92m	16.52m	14.83m	
	First	3.16s	2.69s	4.36s	2.38s	2.87s	
	Last	174.58s	141.18s	108.63s	83.86s	120.67s	
Corvus_2	Total	19.59m	16.99m	13.68m	12.79m	12.03m	
	First	7.09s	3.86s	3.07s	3.10s	3.47s	
	Last	126.65s	160.10s	80.13s	108.79s	83.36s	

6.5 Discussion

Any participant answering all 60 Raven's SPM items in three minutes raises substantial concerns about the legitimacy of the test. A second reason for concern is that the relative change in time taken per test item (the gradients in Figure 6.3, or the times taken for Raven's SPM's final test item in Table 6.2) over the duration of each session for Raven's SPM is unexpected for a measure of fluid intelligence

with incremental difficulty. Difficult test items should on average take participants more time to complete, as is the case for the first sessions of Raven's SPM and every session of Corvus_1 and Corvus_2, but is not the case for the final sessions of Raven's SPM. This might suggest that participants were answering randomly, however their Spearman's ρ was maintained above two thirds, suggesting instead that participants were answering the test by rote memory. This suggests that Raven's SPM ceased acting as a valid measure of fluid intelligence.

Even though Raven's SPM is a test that was very well validated and established in non-longitudinal contexts, these results could be a serious problem for any longitudinal study using Raven's SPM and other traditional tests to measure change in fluid intelligence, with intervals of less than 14 years (Salthouse et al. 2004), and may also be an issue for testing other domains. Validity testing could be completed specifically for repeat testing, rather than cohorts using standard 'off the shelf' psychometric measures that were never designed for longitudinal contexts.

The fact that the learning effect in terms of time taken for Corvus was shown to be so small after the first session, suggests that within-participant change should be easily detectible when using the Corvus paradigm, though a second study including the participants likely to be undergoing such changes would be required to confirm this hypothesis, using the results from this study as normative data.

While on average there was a small improvement in score on repeat measurement, as the rank order was maintained for Raven's SPM and Corvus_2, this is not an issue, as the participant's relative score was unchanged. Corvus_1, however, has a much lower Spearman's ρ than the other two branches (Figure 6.1). As it is likely that on repeat measurement, Raven's SPM and Corvus_2 measure different cognitive features, and Corvus_1 was designed to be positioned between the two, it

is possible that Corvus_1's lower Spearman's ρ is due to being a mixed measure, i.e. it is relatively inconsistent in what it is measuring. This raises a potential issue that I have not seen addressed in the literature; that parallel forms can be too similar. Talking to participants post study corroborated this insight qualitatively; participants who were in the Corvus_1 branch, referred to test items that came at the same time but in different sessions as the same item, while participants in the Corvus_2 branch did not.

The difference in time between first session and subsequent sessions in Corvus_2, seen in Figure 6.3 may be explained by system learning effect and its lack of presence in Corvus_1 could be due to the smaller sample size, or the fact that Corvus_1 appears to be a mixed measure.

At about item 49 of Corvus_2, a sharp change occurs in time taken. While a similar change also occurs in Raven's SPM and Corvus_1, that change is much smoother. This is likely due to a much smoother transition to Logic Gates from distributions of three in Raven's SPM, than in Corvus_2 (and Corvus_1 emulates Raven's SPM much more closely). In other words at that point Corvus_2 deviated from J. Raven's design principle of overlapping test item design (Carpenter et al. 1990). It may be possible to ameliorate this through using the Attributes designed for Logic Gates with distribution–of–three.

Chapter 7

Mouseover Study

7.1 Introduction

One of the main criticisms of Raven's SPM is that it also correlates with working memory, rather than being a 'pure' measure of fluid intelligence. Corvus provides a mouseover feature that was added to the generator due to its perceived potential to ameliorate this issue – and because participants liked it. However this feature, being impossible to implement outside of computer-based testing contexts, is not analogous with traditional paper-based Raven's-like tests, and as such has the potential to impede direct comparisons between them.

A mouseover feature is an event that occurs when a user's mouse moves over an object of interest. Most frequently these are used to trigger displaying 'tool-tips', which are generally pieces of information that are only visible while the user's mouse is positioned over the item they are linked to.

The primary aim of this study is to investigate the effect of Corvus's mouseover feature on test results.

Three secondary goals have also been included in this study, the first of which is to investigate how the mouseover feature might interact with participants' working memory. This was primarily done as an investigation of how the feature interacts with its goal of ameliorating the criticism of Raven's SPM's correlation with working memory, however working memory is also of interest independently as it is sometimes used as a measure of learning ability (Guinagh 1971).

The second is to investigate the effect of the mouseover feature on mouse tracking. In most fluid intelligence tests tracking mouse movement does not accomplish much, as mouse movement rarely benefits participants while they are working on a difficult problem. As a result participants can delay moving their mouse until they have reached a decision, but mouse tracking requires mouse movement to begin early in the decision making process in order to gain data on it (Lins, Schöner, Lins & Sch 2017). However, it is hoped that the mouseover feature will be sufficient to encourage participants to provide the study with additional information via their mouse movements.

The third and final secondary goal is to compare two means of scoring digit span tests; the traditional method, and a method based on measuring the differences between the participant's submissions and the target sequence, using the Damerau-Levenshtein (DL) distance, which is utilised in non-linguistic spell checkers to quantitatively measure the similarity of words. This was done as the traditional method has a significant flaw in that it is too granular — categorising the majority of participants using the five integers; 7 ± 2 (Miller 1956), which results in information loss. Which is particularly a problem for small studies such as this one.

7.2 Design

After acquiring descriptive data, the study asked participants to complete the Reverse Digit Span test before taking three versions of ten item tests sequentially. These ten item tests are generated by Corvus, and participants are systematically assigned to the different orders in which they could take the three tests according to a latin square design. Corvus was enhanced for this study in order to track and record mouse movement, though in a limited way.

The three ten item tests were described to participants during the study as follows;

1. The mouseover functionality is **active** and the options are otherwise **visible**
2. The mouseover functionality is **deactive** and the options are otherwise **visible**
3. The mouseover functionality is **active** and the options are otherwise **hidden**

In essence the three tests differed from each other in how they interacted with the mouse. Test 1 interacted the same way as Corvus has done in the previous two studies. This chapter will sometimes refers to test 1 as 'tool-tip', as its functionality is similar to that of traditional tool-tips. Test 2 had no mouseover interaction and as such is closer to the traditional paper based tests looked at in Chapter 3, Focused Literature Review: Comparison of Established Tests. Test 2 is referred to as the control. Test 3 required participants to move their mouse over each option in order to be able to see it. Test 3 is sometimes referred to as the 'hidden' test.

Although the three tests were randomly generated, they were done so according to the same settings and parameters.

Participant descriptive data included gender, age, level of education and mental health. Education was recorded on a scale of one to five, ranging from the participant's highest level of qualification being GCSEs (1) to having completed a PhD (4). Five indicated a preference not to say, although no participants chose to do so in this study. Mental Health was measured using MHI-5.

7.3 Materials and Methods

7.3.1 Reverse Digit Span

I programmed the reverse digit span test in JavaScript and D3, similarly to Corvus. Each test item displays a target sequence of numbers, one at a time — fading in and out at the same location, the participant is asked to remember the target sequence and then type it in, in the reverse order. The test starts with two practice items, which differ from the standard items in that the borders of the test are coloured green rather than black, the participant can ask the test to repeat the sequence, incorrect elements of responses are highlighted in red, and incorrect answers cannot be submitted.

The keypad is deactivated while the sequence is presented, which is indicated to the participant by it being "greyed-out", and becoming black as soon as the sequence is finished. The keypad provides the numbers zero to nine, in a standard number pad layout, a delete button and a submit button. Every button press is recorded, including inputed characters that are later deleted, as is the final submission. The sequences generated for each test item do not repeat the same digit within five elements, and as the test progresses the sequences get longer —

starting at two and three during the practise, then ranging from three to eleven for the standard test items.

Traditional design would set the test cut–off point after the participant has consecutively answered two items incorrectly (Woods et al. 2011), although there is significant variation between tests, and little documentation of this. All participants in this study answered items at least up to the threshold suggested by Woods et al, if not more, though exactly how many more varies. Initially the code was set to end in the traditional way, given a minimum of four items, but after inspecting the data from the first five participants, I altered the test in two ways. Firstly, the sequence of numbers after the practice sequences was extended from 3,4,5,6,7,8,9,10,11 to 3,4,4,5,5,6,6,7,7,8,8,9,9,10,11. Secondly, the cut–off point was delayed from the traditional rule, to five wrong responses in the last seven items. As the data can be trimmed artificially to fit traditional scoring for analysis purposes, these two changes did not impact the traditional score, and also considerably improved the data with which to test the new scoring method. Five participant test results from pilot testing the Reverse Digit Span software were included in some analysis where only data from the Reverse Digit Span test was required. These pilot test participants completed the full initial sequence from three to eleven, with no suspension.

The early cut–off mechanism is used as it is possible that for some participants test items beyond a certain length provide no information that is useful to the specific aims of this study, and because at least some participants find the test unpleasant or challenging.

Participants are scored using both a traditional scoring method (their highest score before making two consecutive mistakes) and also a novel method using the DL distance between each of their submissions and the correct answer, divided by

the length of the target sequence (which can then be inverted so that high scores indicate a better score and meaning that participant scores are not benefited if the test cuts out early).

Conway et al. (2005) investigated the benefits of partial scoring vs all or nothing scoring, as used by Woods et al. (2011), and found partial scoring to be superior. DL's use in scoring these tests is a logical extrapolation of that finding.

The Reverse Digit Span test fades each element of each sequence in and out over 1500ms. The beginning and end of each sequence is signalled by the animated appearance and disappearance (or opening and closing) of the whitespace the elements of the sequence are displayed in, and the keypad is disabled until the sequence had finished. Before beginning the display of each sequence the test waits 600ms after the submission of the previous sequence.

Damerau-Levenshtein

DL allows us to quantify the magnitude of a participant's error. For example, if the target sequence was '12345', and the participant gave the answer '12344', the traditional scoring method would consider them just as wrong as had they had answered '09876' instead. DL allows us to differentiate between those two responses.

DL distance is the 'edit distance' or quantitative similarity between two lists or sequences. Edit distance is the minimum number of operations required to get from one sequence to the other. For Damerau-Levenshtein the permitted operations are insertion, deletion, substitution and transposition.

For example, given the sequences $\{1, 2, 3\}$, and $\{2, 1\}$, we can get from the first to the second by using transposition on the first two elements, followed by deleting

the three at the end; i.e. these two example sequences have a distance of two. A distance of zero indicates that the two sequences are identical.

Levenshtein distance (sans Damerau) is the same, excepting that it does not allow for transposition — the example above would have a Levenshtein distance of three, as the transposition would have to be replaced with two substitutions. Damerau-Levenshtein is preferred in this context as if the correct answer was $\{1, 2\}$ it measures the response $\{2, 1\}$ as better than $\{8, 9\}$, while Levenshtein distance would result in them being considered as bad as each other. This means that the participant's response garners a lower distance from the correct result for remembering the right digits in the wrong order, than for not remembering the correct digits at all. It is not clear which of the two algorithms is optimal for measuring working memory; comparison between the two measures could be an interesting avenue for future work as the difference in distance has the potential to interact with any difference between serial recall and free recall. Both of these measures of distance are often used in non-linguistic spell-checkers. For the purposes of this study DL distance was used.

The DL distance is then divided by the length of the target sequence, added to 1, and then inverted to acquire the test item's score. DL distance is first divided by the length of the target sequence as it means that a participant making one mistake in a sequence of length eleven scores higher than if they made one mistake in a sequence of length three. Dividing by the length of the sequence prevents loading the test items by difficulty, which makes sense as the test items all measure the same underlying ability (Conway et al. 2005) and has been shown to be superior in terms predicting of achievement (Clair-Thompson & Sykes 2010), this is despite a specific tradition in working memory (see Ebbinghaus 1897) that goes counter

to wider psychometric practices. The addition of one is done to prevent division by zero when inverting. The inversion is implemented so that higher scores are better, and that answering additional test items provides opportunity to score higher overall (when test item results are summed).

A score of one on a test item is a perfect answer, and scores that tend towards zero indicate submissions that are increasingly divergent from the sequence. Scoring zero on an answered test item is impossible, but unanswered items can be either be assigned a score of zero, or excluded from the calculation.

The score for the test as a whole is the sum of the scores for each test item, divided by the number of test items answered. This step of dividing by the number of test items answered is necessary in the particular case of this study as participants had different halt criteria. It may also be helpful in preventing skewed results arising from halt criteria that was chosen arbitrarily, or due to concerns about participant enjoyment or retention, rather than on the basis of accurate measurement. Further investigation of optimal halt criteria is needed for this scoring method.

$$\text{Experimental Test Score} = \sum_{i=1}^{i=\text{test length}} \frac{1}{\frac{\text{DL distance}_i}{\text{Target Sequence Length}_i} + 1}$$

$$\text{Traditional Test Score} = \text{Test length (given particular end criteria)}$$

Digit span tests have previously used much simpler scoring methods; generally using how long the participants managed to not trigger their test's cut-off mechanism (Conway et al. 2005; Woods et al. 2011), which is easier for human invigilators to grade on the fly. However it was felt that the method of scoring used here, and that is enabled by the use of computer-based testing, was more intuitive, in that

the participant's responses are essentially spell-checked against the target sequence, because it might provide a more sensitive measure, and because the granularity of the traditional method was an issue for the small size of this study. The test's cut-off point was chosen to always go longer than the cut off used in traditional tests such as Woods et al. (2011), so that this measure could be compared with traditional measures, though this was not one of the primary aims of the study.

7.3.2 Corvus Generated Tests

All three versions of the Raven's-like tests are independently randomly generated with Corvus, using the same functions as were used to generate the Corvus_2 branch in the learning effects study, though the test item definitions used for all three branches were limited to ten items selected using every sixth item from the definitions used in the previous study. Albeit with some variance from the six item leaps to avoid odd or unusual items.

The three versions only differ from each other in regards to how they interact with mouseovers: the first uses the mouseover feature as used in the other two studies involving Corvus, where a participant moving their mouse over any option causes that option to appear in the matrix, as if filling in the missing element. The second has no mouseover feature, and in this is more similar to Raven's SPM. For the third version, the exact nature of the options are hidden by default, though their location and number is not, and when participants move their mouse over the location of the item, it is revealed — and also placed in the matrix, in the same way that occurs in version 1. The details of the option are then hidden again the moment the mouse is moved away.

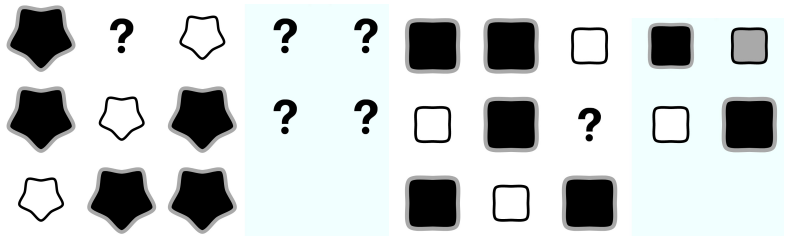


Figure 7.1: From left to right, respectively; examples of Hidden and Control versions. The question marks concealing each available option (in the light blue zone on the right hand side of each item) in the Hidden version are removed, causing the option to become visible while the participant positions their mouseover that option.

For all three versions a new feature has been added to Corvus, in that it now records (specifically and only) whenever the participant's mouse is moved over or away from any of the options (not including the button allowing participants to skip test items). These are data that Corvus had previously processed in that it was responding to mouse movement as part of the mouseover feature, but now it also records that information.

7.4 Results

There were 15 participants consisting of 7 men and 8 women, whose aged ranged from 18 – 66 years old, with a mean age of 40. They had normal or corrected to normal vision and had access to the internet via a computer with a mouse or trackpad. Their mean MHI-5 score was 67.07, and mean education was 2.04 (on the same scale used in the previous studies). The correlation between education and test score had a mean of $r = 0.24$.

Two anomalous results (one participant, and one individual test item from another participant) were excluded from all analysis involving test time.

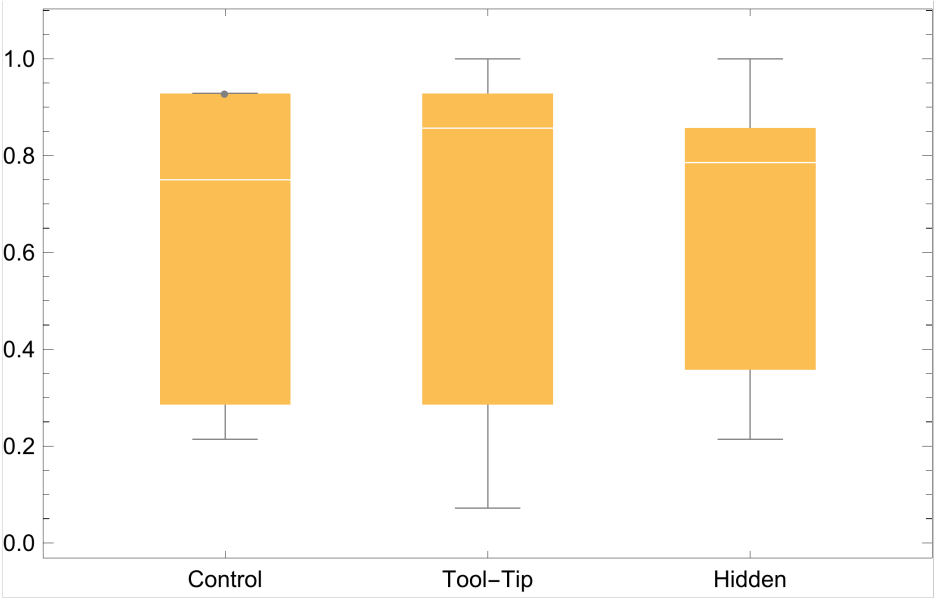


Figure 7.2: Graph comparing mean item test scores, interquartile range, and range from the three branches of Corvus.

7.4.1 Impact of Standard Mouseover

No statistically significant change was detected between the control and mouseover versions for test scores, or time taken (see Figures 7.2 and 7.3 respectively).

7.4.2 Working Memory

No statistically significant correlation was found between any of the three versions of Corvus, and either measure for the reverse digit span test (All three versions of Corvus are plotted against Reverse Digit Span test using DL distance score in Figure 7.4).

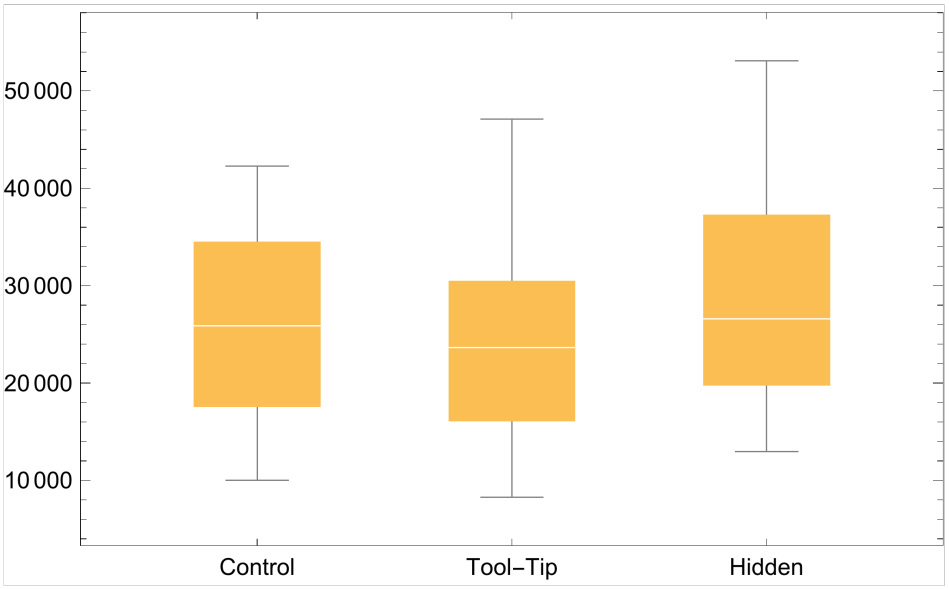


Figure 7.3: Graph comparing mean item test time, interquartile range, and range in ms from the three branches of Corvus.

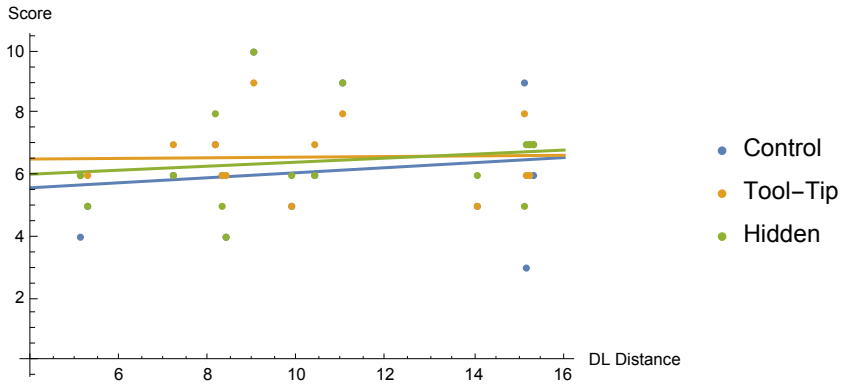


Figure 7.4: Graph comparing DL distance and scores from the three branches of Corvus.

7.4.3 Mouse Tracking

Although no statistically significant difference was detected between the control and standard mouseover versions, a substantial difference was detected between each of those and the third version, where the options were hidden unless the mouse was positioned over them.

Mouse movement between options was dramatically increased as can be seen in Figure 7.5.

Due to idiosyncrasies with how Corvus internally handles which option is which, and their ordering; there are two ways of recording which option the mouse was moved over, and which option was clicked on (before and after shuffling them). The mouseover feature recorded both of these methods, while the code that recorded which option was selected by the participant only recorded one (though it has always recorded the random number generator seed that was used to construct the test item, so although some information was not recorded, means of reconstructing it manually was). This resulted in some additional steps required in the analysis. The mouse tracking used in this study was not detailed, and when coding it I omitted tracking when a mouse was moved over, or away from the button to skip test items. The exact position of the mouse is not recorded — though the capability to do so is an obvious potential addition to Corvus in the future. These issues highlighted means by which the data recording code could be improved.

7.4.4 Reverse Digit Span Test Scoring

The two methods of scoring the Reverse Digit Span test correlated well ($r = 0.79$), as can be seen in Figure 7.6. The data included in Figure 7.6 includes data from

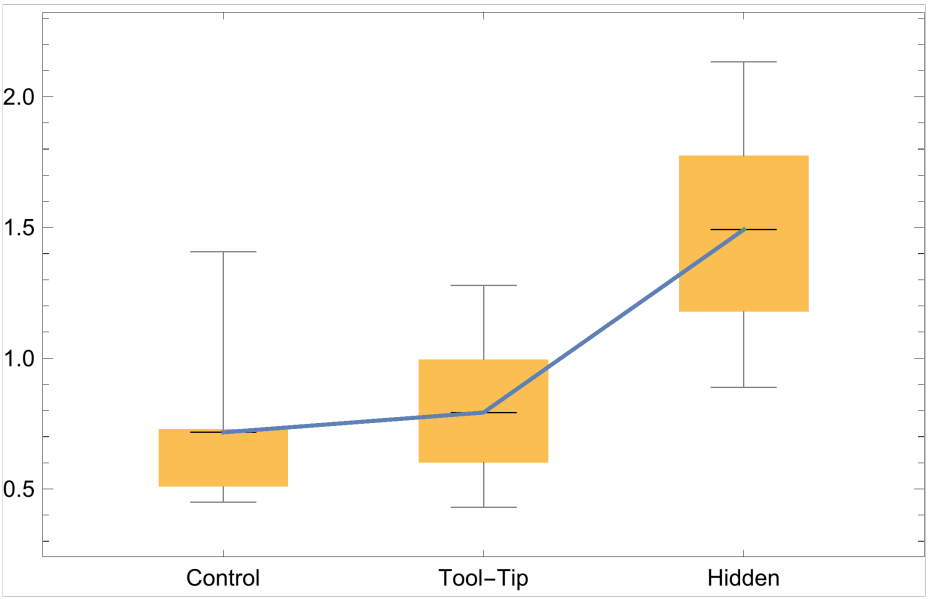


Figure 7.5: Graph comparing mean mouse movement, interquartile range, and range between the three branches of Corvus.

five participants from pilot testing the Reverse Digit Span test code — all five of which lie fairly centrally in the graph.

For the traditional method only three score categories included more than two participants (3, 4, or 5), while the new scoring method provided a nearly continuous scoring range.

7.5 Discussion

7.5.1 Impact of Standard Mouseover

Though small, this study suggests that the mouseover feature that is present in the Corvus generated tests looked at in the previous chapters does not invalidate

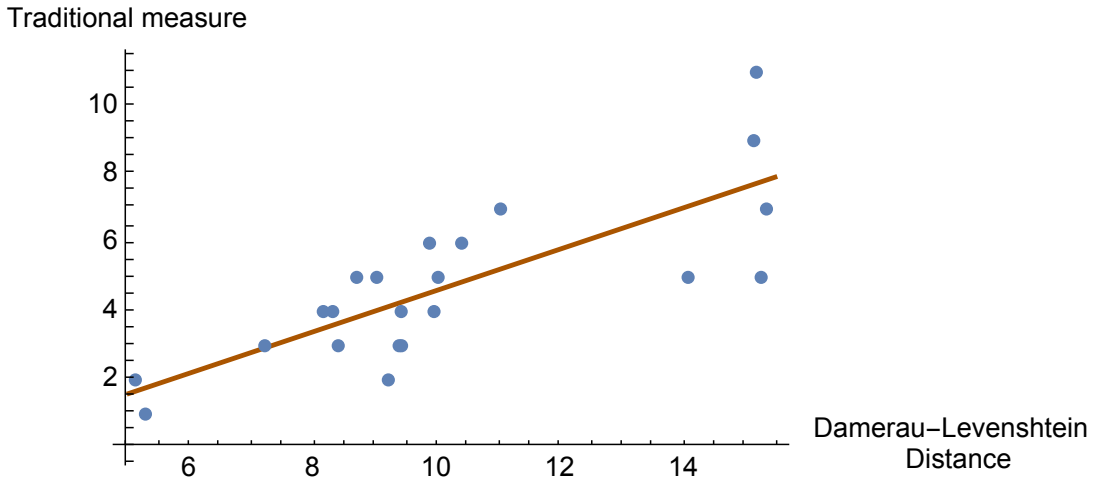


Figure 7.6: Graph plotting DL distance against traditional measure of working memory.

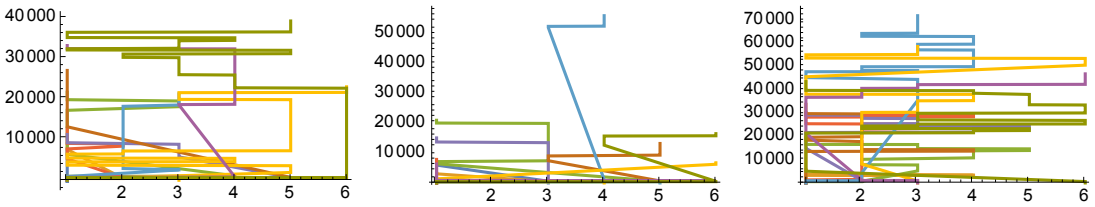


Figure 7.7: Graphs showing an example of how mouse movement changes between ten item tests.

comparisons between those tests and tests without similar mouseover features, such as Raven's SPMs.

7.5.2 Working Memory

The lack of any correlation between working memory and the Raven's-like tests runs was unexpected. Especially as the traditional measure of a reverse digit span test is a well established measure of working memory, which is in turn well established as correlating with Raven's SPMs. It is supposed that this lack of correlation in the data is primarily due to the study being underpowered for this purpose.

7.5.3 Mouse Tracking

The hidden version of Corvus produced promising results, though further studies would be required to investigate its potential uses. However, doubling the information gain from mouse tracking, combined with little to no impact on test time, provides strong motivation for inclusion in future studies.

As the options are only visible in the hidden version when the participant has their mouse positioned over that option, there is a more obvious link between mouse tracking and eye-tracking. Research into eye-tracking has shown that up to 41% of variance in Raven's Advanced Progressive Matrices could be attributed to strategy that is detectible with eye-tracking data (Hayes 2017), this suggests that it may be possible to use eye-tracking or mouse tracking data to mitigate test variance from strategy in order to score participants more accurately; a participant that scores well despite employing poor strategies could provide interesting information about

that participant. A similar idea is explored for eye-tracking in Hayes et al. (2015). Example mouse tracking data from the study can be seen in Figure 7.7.

Epidemiologists invest a great deal of effort into designing test batteries, and have multiple considerations when doing so. One of the chief considerations is participant retention and test duration. One of the solutions to this is to use CAT which increases the information gain per item, allowing a shorter test to gather a similar amount of information as the full length version. Another solution is to use lower resolution tests than would be used in standard psychometric settings. This works because epidemiologists deal with very large study populations, so long as the shorter test is traded off against individual-level test reliability, rather than test validity. Similarly mouse tracking allows investigators to gather more information with each test item, and as such provides a new potential solution that could be used in concert with the other methods. Further investigation would be required to see if this enabled fewer administering fewer test items, but regardless of the length of the test, it does increase the efficiency of the test, as more useful data is gathered in the same amount of time.

While no statistically significant difference was detected between the control and standard mouseover versions, a larger sample size may have found one. Nonetheless this result suggests that any difference would be small.

7.5.4 Reverse Digit Span Test Scoring

Scores via the traditional method is lower than the 7 ± 2 Miller (1956) would suggest, which may be due to a combination of it being a reverse span — adding an

extra layer of cognition on top of forward spans, the presentation of the sequence being visual rather than auditory, and the test being computer based.

The new scoring method does solve some of the traditional method's problems. However the new scoring method may bring with it its own issues. Foremost that it wants participants to complete as many test items as possible and that this is not a very pleasant test to do. That said, it may be possible to mitigate the increase in test length via CAT.

Establishing the best method of scoring would require a study specifically looking at that topic, which is unfortunately beyond the scope of this thesis.

7.5.5 Summary

The primary motivation behind this study were concerns about the validity of comparisons between Corvus and traditional tests done in the previous two studies due to the mouseover feature. No evidence supporting those concerns was found, as the differences between the control and 'tool–tip' versions of Corvus are small for both test score and test time.

Although this study was underpowered to achieve its secondary goals fully, it was able to produce promising results for mouse tracking and the alternative method for scoring digit spans suggesting that further investigation could be worthwhile. Also though the study did not find any interactions between working memory and the mouseover feature, this result could also be considered as additional evidence pointing towards the lack of negative impact by the mouseover feature on test comparability, and thus provide supporting evidence for the study's primary finding.

Chapter 8

Synopsis

8.1 Introduction

Cohort studies require numerous and potentially frequent retests when looking for behavioural and psychological change in participants. The construction and validation of tests requires a substantial amount of repetitive work, which is expensive to do manually. As a practical result, frequent retests require either retesting using identical forms, or automated test generation.

The first part of this thesis investigates automated test generation, including other established Raven's-like tests, the second part details the development of Corvus. Lastly the final part starts to validate Corvus, and begin using it to explore some interesting questions. These questions are not fully explored here, and doing so is left to future work — particularly the secondary goals investigated in chapter 7.

There are a number of barriers inhibiting fully automated test generation from being deployed. One such difficulty is item difficulty estimation, on which work has

begun elsewhere (e.g. Chan 2018; Freund et al. 2008; Matzen et al. 2010; Ragni et al. 2011)[1], and is essential for Item Response Theory and adaptive testing. Another potential issue is in investigating learning effects, and making sure that retests are comparable to other retests, for which some work has been done on investigating with higher numbers of retests (e.g. Bartels et al. 2010) and is well explored for less than three retests (Scharfen et al. 2018). However, an area of interest that is currently lacking in research, is that of investigating if and how retests retain validity (Scharfen et al. 2018) — which directed the focus for one of the studies completed as part of this thesis, and is detailed in Chapter 6.

8.2 Corvus

The test generator, Corvus, was designed with the intent of being able to produce tests that had a controllable degree of similarity or emulation to a variety of established paper-based Raven's-like tests (Chapter 4). Through analysis of the literature, and of the established tests (Chapters 2 and 3, respectively), it was found that Raven's-like option sets were an area that was under-researched, and under-recognised as a subject with significant impact on test items, while the intricacies of matrices were relatively well researched.

In fact Meo et al. (2007), is the only paper referenced in my literature review (Chapter 2) that cites the first paper on Clues and Anti-Clues, White and Zammarelli

[1]It is worth noting that this evaluation of the impact on difficulty by the specific properties of test items relate directly to the psychometric terms 'rules' and 'incidentals', and as such they are also outside the scope of this thesis. 'Rules' in that context means relevant features, and 'incidentals' means irrelevant or cosmetic features. A test item property that is found to have impact on test item difficulty is a 'rule' (in this sense), and something that has no impact is an 'incidental'. To avoid confusion these terms are avoided throughout the rest of this thesis, though it is worth noting that if any property of a test item is a 'rule', then the Rules (in the sense of the word that this thesis uses) are perhaps the most obvious candidates.

(1981), despite the fact that many of the other papers would have benefited from it (e.g. Beckmann 2008).

Similarly, though not to the same extent, only a small percentage of papers mentioned Logic Gates by name. Further to that omission, of the papers that detailed the types of Rules provided for by any automatic item generator they had developed or used, a surprisingly large number silently left out one or more of the Logic Gates.

Corvus was found to be capable of generating tests to specification, generating valid tests at baseline (Chapter 5), and no evidence against its validity on repeat testing was found (Chapter 6).

8.3 Studies

The learning effects study in Chapter 6 had three primary findings, which taken together provide evidence that increasingly repeated tests of Raven's SPM (as an identical form) are decreasingly valid tests of fluid intelligence, and have an increased memory component. The findings were firstly that as participants took more retests the amount of time participants took for each test item had a decreasing correlation with item difficulty, and tended towards a constant; secondly that the time participants took to do all 60 items dropped by more than a sixth from over half an hour to less than three and a half minutes — or in other words less than three and a half seconds per item; and finally that despite these changes, participant rank order was maintained over retests. The implication of these results is that sufficiently unique forms are increasingly necessary for maintaining the validity of repeated retests, albeit with short intervals. If these findings are maintained over

longer time periods, then this requirement is potentially the case for the validity of retests in general.

Following the validation (Chapter 5) and learning effects studies (Chapter 6), a third study was conducted with the primary aim of addressing an uncertainty involving the mouseover feature, which had the potential to interfere with comparisons between Corvus and traditional tests (Chapter 7). The study found no statistically significant impact on test score or test time by the mouseover feature, indicating that if there is an effect, then it is small. This suggests that the mouseover feature did not interfere with comparisons done between Corvus and other tests in the previous studies.

Additionally a number of secondary goals were also attached to this third study. Of particular note for the context of cohorts (which often conduct cognitive studies online due to the economy of scale) were the results from mouse tracking. The indication from these results was that additional information could be extracted from participants in virtually the same amount of time, through forcing participants to move their mouse over the screen objects they wished to look at; effectively providing eye-tracking for online contexts without specialised hardware.

The study's other secondary goal was to improve scoring from digit spans. The new scoring method had potential in that while it might introduce its own issues, it did address the problems with traditional scoring methods. Both of these subsidiary goals would benefit from further investigation, but a promising start has been made here.

8.4 Future Research

Perhaps the next logical step for future research inspired by my work, would be to conduct a similar study as done in Chapter 6, Learning Effects Study but with significantly longer intervals — perhaps six months to a year between retests, rather than days. I suspect a similar effect would eventually occur as that found in the shorter study here. Particularly as cohorts such as the Caerphilly study had obvious retest effects with intervals of five years (Caerphilly Prospective Study, personal communication, 2018) despite its ageing population. Though these retest effects might take more sessions to occur to the same degree as they did in this learning effects study, and such a study may need to take the Flynn effect into consideration. It would also be interesting to run a similar study on a different cognitive domain, however that would likely require the development of another test generator.

Another interesting investigation arising out of Chapter 6, would be to explore exactly when and how test forms can get too similar.

It could also be fruitful to investigate how and when incorrect answers correlate with intelligence and how incorporating that information back into scoring methods might impact tests. Particularly in the light of Cattell's Culture Fair's incorrectly marked items still correlating with fluid intelligence. In other words if and how non-dichotomous scoring effects intelligence tests, and how that might interact with adaptive testing.

While the results of the third study were positive with regards to mouse tracking and the new scoring method for digit spans, nonetheless the study was not designed primarily to investigate these topics, and little relating to these secondary goals can be confidently stated on the basis of this study. What can be concluded about

them from the study however, is that both avenues of research have potential and warrant further investigation.

While Corvus was sufficient for the purposes of this thesis, as a test generator Corvus still has room for improvement. In particular, customising its settings currently requires knowledge of JavaScript, so it would benefit from a graphical user interface for investigators, which would be routine programming, but was not relevant to the goals of the thesis. Beyond that, there is always space for adding new features to any tool like this. Corvus has been made available under an open source licence so that others can use its code and contribute to its future development should they wish (see Appendices B, page 194).

It is hoped that my work in analysing traditional tests (Chapter 3), and on the development of Corvus (Chapter 4) will continue to contribute and inspire research into understanding and constructing Raven's-like tests, and will encourage a more detailed examination of option sets, and in particular their effect on item difficulty estimation, as well as a wider recognition of the work originating with White and Zammarelli (1981).

The crux of this thesis is that, despite searching for it, no evidence was found indicating that Corvus failed to maintain validity on repeat testing. Corvus's validity on repeat testing is an important advantage over established tests, such as Raven's SPM. Hopefully Corvus, or something inspired by it and the research presented here, will be developed sufficiently for easy and widespread use by cohort studies, so that they can frequently retest fluid intelligence, while maintaining validity.

References

Anastasi, A. (1981). Coaching, Test Sophistication, and Developed Abilities. *American Psychologist*, *36*(10), 1086–1093. doi: 10.1037/0003-066X.36.10.1086

Arendasy, M. (2005). Automatic Generation of Rasch-Calibrated Items: Figural Matrices Test GEOM and Endless-Loops Test EC. *International Journal of Testing*, *5*(3), 197–224.

Arendasy, M. & Sommer, M. (2005). The Effect of Different Types of Perceptual Manipulations on the Dimensionality of Automatically Generated Figural Matrices. *Intelligence*, *33*(3), 307–324. doi: 10.1016/j.intell.2005.02.002

Arendasy, M. & Sommer, M. (2012). Gender Differences in Figural Matrices: The Moderating Role of Item Design Features. *Intelligence*, *40*(6), 584–597. doi: 10.1016/j.intell.2012.08.003

Arendasy, M. & Sommer, M. (2013). Reducing Response Elimination Strategies Enhances the Construct Validity of Figural Matrices. *Intelligence*, *41*(4), 234–243. doi: 10.1016/j.intell.2013.03.006

Arendasy, M. & Sommer, M. (2017). Reducing the Effect Size of the Retest Effect: Examining Different Approaches. *Intelligence*, *62*, 89–98. doi: 10.1016/j.intell.2017.03.003

Bachoud-Lévi, A.-C., Maison, P., Bartolomeo, P., Bachoud, A. . C., Lévi, ., Boissé, M. . F., ... Peschanski, M. (2001). Patients with Early HD Retest Effects and Cognitive Decline in Longitudinal Follow-up of Retest Effects and Cognitive Decline in Longitudinal Follow-up of Patients with Early HD. *Neurology*, *56*, 1052–1058. doi: 10.1212/WNL.56.8.1052

Bartels, C., Wegrzyn, M., Wiedl, A., Ackermann, V. & Ehrenreich, H. (2010). Practice Effects in Healthy Adults: A Longitudinal Study on Frequent Repetitive Cognitive Testing. *BMC Neuroscience*, *11*, 118. doi: 1471-2202-11-118 [pii]10.1186/1471-2202-11-118

Beckmann, B. M. E. (2008). *Reasoning Ability: Rule-based test construction of a figural analogy test* (Unpublished doctoral dissertation).

Bors, D. A. & Vigneau, F. (2001). The effect of practice on Raven's Advanced Progressive Matrices. *Learning and Individual Differences*, *13*(4), 291–312. doi: 10.1016/S1041-6080(03)00015-3

Calamia, M., Markon, K. & Tranel, D. (2012). Scoring Higher the Second Time Around: Meta-Analyses of Practice Effects in Neuropsychological Assessment. *Clinical Neuropsychologist*, *26*(4), 543–570. doi: 10.1080/13854046.2012.680913

Carlesimo, G., Caltagirone, C., Gainotti, G., Fadda, L., Gallassi, R., Lorusso, S., ... Parnetti, L. (1996). The Mental Deterioration Battery: Normative Data, Diagnostic Reliability and Qualitative Analyses of Cognitive Impairment. *European Neurology*, *36*, 378–384. doi: 10.15713/ins.mmj.3

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

Cassady, J. & Gridley, B. (2005). The Effects of Online Formative and Summative Assessment on Test Anxiety and Performance. *Journal of Technology, Learning, and Assessment*, *4*(1).

Catron, D. W. & Thompson, C. C. (1979). Testretest gains in WAIS scores after four retest intervals. *Journal of Clinical Psychology*, *35*(2), 352–357. doi: 10.1002/1097-4679(197904)35:2¡352::AID-JCLP2270350226¿3.0.CO;2-2

Cattell, R. B. & Cattell, A. K. S. (1961). *Culture Fair Intelligence Tests*. Hogrefe Ltd.

Cervilla, J., Prince, M., Joels, S., Lovestone, S. & Mann, A. (2004, aug). Premorbid cognitive testing predicts the onset of dementia and Alzheimer's disease better than and independently of APOE genotype. *Journal of neurology, neurosurgery, and psychiatry*, *75*(8), 1100–6. doi: 10.1136/jnnp.2003.028076

Chan, Y. W. F. (2018). *Development of Matrices Abstract Reasoning Items to Assess Fluid Intelligence* (Unpublished doctoral dissertation). University of Cambridge.

Chang, H. (2004). *Inventing temperature: Measurement and scientific progress*. Oxford University Press.

Clair-Thompson, H. S. & Sykes, S. (2010). Scoring methods and the predictive ability of working memory tasks. *Behavior Research Methods*, *42*(4), 969–975. doi: 10.3758/BRM.42.4.969

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhem, O. & Engle, R. W. (2005). Working memory span tasks : A methodological review and user ' s guide. *Psychonomic Bulletin & Review*, *12*(5), 769–786. doi: 10.3758/BF03196772

Crystal, D. (1995). *The Cambridge Encyclopedia of the English Language*.

Daniele, A., Albanese, A., Contarino, M. F., Zinzi, P., Barbier, A., Gasparini, F., ... Scerrati, M. (2003). Cognitive and behavioural effects of chronic stimulation of the subthalamic nucleus in patients with Parkinson's disease. *J Neurol Neurosurg Psychiatry*, *74*(2), 175–182.

de Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*. New York: The Guildford Press.

Dimitriou, D., Le Cornu Knight, F. & Milton, P. (2015). The Role of Environmental Factors on Sleep Patterns and School Performance in Adolescents. *Frontiers in Psychology*, *6*(December), 1–9. doi: 10.3389/fpsyg.2015.01717

Dweck, C. S. (2000). *Self-Theories*. New York: Psychology Press.

Ebbinghaus, H. (1897). Uber eine neue Methode zur Prufung geistiger Fahigkeiten und ihre Anwendung bei Schulkindern. *Zeitschrift für die Psychologie*, *13*, 401–459.

Embretson, S. E. (1998). A Cognitive Design System Approach to Generating Valid Tests: Application to Abstract Reasoning. *Psychological Methods*, *3*(3), 380–396. doi: 10.1037/1082-989X.3.3.380

Embretson, S. E. (1999). Generating Items During Testing: Psychometric Issues and Models. *Psychometrika*, *64*(4), 407–433. doi: 10.1007/BF02294564

Estrada, E., Ferrer, E., Abad, F. J., Román, F. J. & Colom, R. (2015). A General Factor of Intelligence Fails to Account for Changes in Tests' Scores after Cognitive Practice: A Longitudinal Multi-Group Latent-Variable Study. *Intelligence*, *50*, 93–99. doi: 10.1016/j.intell.2015.02.004

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin*, *95*(1), 29–51. doi: 10.1037/0033-2909.95.1.29

Flynn, J. R. (1987). Massive IQ Gains in 14 Nations: What IQ Tests Really Measure. *Psychological Bulletin*, *101*(2), 171–191. doi: 10.1037/0033-2909.101.2.171

Flynn, J. R. (1999). Searching for justice: The discovery of IQ gains over time. *American Psychologist*, *54*(1), 5–20. doi: 10.1037/0003-066X.54.1.5

Freund, P. A., Hofer, S. & Holling, H. (2008). Explaining and controlling for the psychometric properties of computer-generated figural matrix items. *Applied Psychological Measurement*, *32*(3), 195–210. doi: 10.1177/0146621607306972

Freund, P. A. & Holling, H. (2011a). How to get really smart: Modeling retest and training effects in ability testing using computer-generated figural matrix items. *Intelligence*, *39*(4), 233–243. doi: 10.1016/j.intell.2011.02.009

Freund, P. A. & Holling, H. (2011b). Retest effects in matrix test performance: Differential impact of predictors at different hierarchy levels in an educational setting. *Learning and Individual Differences*, *21*(5), 597–601. doi: 10.1016/j.lindif.2011.07.006

Galasko, D., Abramson, I., Corey-Bloom, J. & Thal, L. J. (1993). Repeated exposure to the Mini-Mental State Examination and the Information-Memory-Concentration Test results in a practice effect in Alzheimer's disease. *Neurology*, *43*(0028-3878), 1559–1563.

Gallacher, J., Bayer, A., Dunstan, F., Yarnell, J., Elwood, P. & Ben-Shlomo, Y. (2009). Can we understand why cognitive function predicts mortality? Results from the Caerphilly Prospective Study (CaPS). *Intelligence*, *37*(6), 535–544. doi: 10.1016/j.intell.2009.02.004

Gallacher, J., Elwood, P. C., Hopkinson, C., Rabbitt, P. M. A., Stollery, B. T., Sweetnam, P. M., ... Huppert, F. A. (1999). Cognitive function in the Caerphilly study: Associations with age, social class, educa-

tion and mood. *European Journal of Epidemiology*, *15*(2), 161–169. doi: 10.1023/A:1007576324313

Geerlings, H., van der Linden, W. J. & Glas, C. A. (2012). Optimal Test Design With Rule-Based Item Generation. *Applied Psychological Measurement*, *37*, 140–161. doi: 10.1177/0146621612468313

Geyer, J., Insel, P., Farzin, F., Sternberg, D., Hardy, J. L., Scanlon, M., ... Weiner, M. W. (2015). Evidence for age-associated cognitive decline from Internet game scores. *Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring*, *1*(2), 260–267. doi: 10.1016/j.dadm.2015.04.002

Giambra, L. M., Arenberg, D., Zonderman, A. B., Kawas, C. & Costa, P. T. (1995). Adult Life Span Changes in Immediate Visual Memory and Verbal Intelligence. *Psychology and Aging*, *10*(1), 123–139. doi: 10.1037/0882-7974.10.1.123

Goldstein, G. & Hersen, M. (Eds.). (2000). *Handbook of Psychological Assessment*. Pergamon.

Guinagh, B. J. (1971). An Experimental Study of Basic Learning Ability and Intelligence in Low-Socioeconomic-Status Children. *Society for Research in Child Development*, *42*(1), 27–36.

Guttman, L. & Schlesinger, I. (1967, oct). Systematic Construction of Distractors for Ability and Achievement Test Items. *Educational and Psychological Measurement*, *27*(3), 569–580. doi: 10.1177/001316446702700301

Harrington, K. D., Dang, C., Lim, Y. Y., Ames, D., Laws, S. M., Pietrzak, R. H., ... Maruff, P. (2018). The effect of preclinical Alzheimer's disease on age-related changes in intelligence in cognitively normal older adults. *Intelligence*, *70*(May), 22–29. doi: 10.1016/j.intell.2018.07.004

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T. & Moriarty Gerrard, M. O. (2007a). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology*, *92*(2), 373–385. doi: 10.1037/0021-9010.92.2.373

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T. & Moriarty Gerrard, M. O. (2007b). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology*, *92*(2), 373–385. doi: 10.1037/0021-9010.92.2.373

Hayes, T. R. (2017). Mechanisms of visual relational reasoning. *Dissertation Abstracts International: Section B: The Sciences and Engineering*, *77*(11-B(E)).

Hayes, T. R., Petrov, A. A. & Sederberg, P. B. (2015). Do We Really Become Smarter When Our Fluid-Intelligence Test Scores Improve? *Intelligence*, *48*, 1–14. doi: 10.1016/j.intell.2014.10.005

Heim, A. W. (1967). *AH4 Group Test of Intelligence*. London: National Foundation for Educational Research.

Hornke, L. F., Küppers, A. & Etzel, S. (2000). Konstruktion und Evaluation eines adaptiven Matrizentests [Design and evaluation of an adaptive matrices test]. *Diagnostica*, *46*(4), 182–188.

Ian, J., Martha, C., John, M., Lawrence, J. & Helen, C. (2014). The Impact of Childhood Intelligence on Later Life: Following Up the Scottish Mental Surveys of 1932 and 1947. *Journal of Personality and Social Psychology*, *86*(January 2004), 1–25.

Isella, V., Atzeni, L., Iurlaro, S., Villa, M. L., Russo, A., Forapani, E., ... Appollonio, I. M. (2003). Assessing clinically relevant cognitive decline: Prelim-

inary data on a new method. *Neurological Sciences*, *24*(4), 236–241. doi: 10.1007/s10072-003-0146-7

Jaeggi, S. M., Buschkuehl, M., Jonides, J. & Perrig, W. J. (2008). Improving Fluid Intelligence with Training on Working Memory. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(19), 6829–6833. doi: 10.1073/pnas.0801268105

Johnson, W., Corley, J., Starr, J. M. & Deary, I. J. (2011, jan). Psychological and Physical Health at Age 70 in the Lothian Birth Cohort 1936: Links With Early Life IQ, SES, and Current Cognitive Function and Neighborhood Environment. *Health psychology : official journal of the Division of Health Psychology, American Psychological Association*, *30*(1), 1–11. doi: 10.1037/a0021834

Joint Committee on Standards for Educational and Psychological Testing. (2014). *Standards for Educational and Psychological Testing* (American Psychological Association, American Educational Research Association & National Council on Measurement in Education, Eds.).

Kaplan, R. M. & Saccuzzao, D. P. (2018). Standardized tests in education, civil service, and the military. In *Psychological testing: Principles, applications and issues* (9th ed., pp. 322–324). Boston: Cengage Learning.

Karen, R., Guilhem, R., Craig W, R., Alain, B., Vanessa, P., Sylvaine, A., ... Ritchie, K. (2014). COGNITO: Computerized Assessment of Information Processing. *J Psychol Psychother*, *4*(4). doi: 10.4172/2161-0487.1000136

Kline, P. (2000). *Handbook of Psychological Testing* (2nd ed.). Routledge.

Kulik, J. A., Kulik, C.-l. C. & Bangert, R. L. (1984). Effects of practice on aptitude and achievement test scores. *American Educational Research Journal*, *21*(2), 435–447. doi: 10.2307/1162453

Květon, P., Jelínek, M., Voboil, D. & Klimusová, H. (2007). Computer-based tests: the impact of test design and problem of equivalency. *Computers in Human Behavior*, *23*(1), 32–51. doi: 10.1016/j.chb.2004.03.034

Levy, R. & Post, F. (1975). The use of an interactive computer terminal in the assessment of cognitive function in elderly psychiatric patients. *Age & Ageing*, *4*(2), 111–115 5p. doi: ageing/4.2.110

Lievens, F., Reeve, C. L. & Heggestad, E. D. (2007). An Examination of Psychometric Bias Due to Retesting on Cognitive Ability Tests in Selection Settings. *Journal of Applied Psychology*, *92*(6), 1672–1682. doi: 10.1037/0021-9010.92.6.1672

Linacre, J. M., Chae, S., Kang, U. & Jeon, E. (2000). Computer Adaptive Testing: A Methodology Whose Time Has Come. *Development of Computerized Middle School Achievement Test [in Korean]*, 1–31. Retrieved from `http://www.cehd.umn.edu/EdPsych/C-BAS-R/Docs/Linacre2000_CAT.pdf`

Lins, J., Schöner, G., Lins, J. & Sch, G. (2017). Mouse Tracking Shows Attraction to Alternative Targets While Grounding Spatial Relations Mouse Tracking Shows Attraction to Alternative Targets. (July).

Martin, G. N., Carlson, N. R. & Buskist, W. (2010). *Psychology* (4th ed.). Pearson Education Limited.

Matton, N., Vautier, S. & Raufaste, É. (2009). Situational effects may account for gain scores in cognitive ability testing: A longitudinal SEM approach. *Intelligence*, *37*(4), 412–421. Retrieved from `http://dx.doi.org/10.1016/j.intell.2009.03.011` doi: 10.1016/j.intell.2009.03.011

Matzen, L. E., Benz, Z. O., Dixon, K. R., Posey, J., Kroger, J. K. & Speed, A. E. (2010). Recreating Raven's: Software for systematically generating large numbers of Raven-like matrix problems with normed properties. *Behavior Research Methods*, *42*(2), 525–541. doi: 10.3758/BRM.42.2.525

McCaffrey, R., Duff, K. & Westervelt, H. (2000). *Practitioner's guide to evaluating change with neuropsychological assessment instruments*. Springer US.

McCallum, S., Bracken, B. & Wasserman, J. (2000). *Essentials of Nonverbal Assessment*. John Wiley & Sons.

Meo, M., Roberts, M. J. & Marucci, F. S. (2007). Element Salience as a Predictor of Item Difficulty for Raven's Progressive Matrices. *Intelligence*, *35*(4), 359–368. doi: 10.1016/j.intell.2006.10.001

Miller, G. A. (1956). The Magical Number Seven, Plus-Or-Minus Two Or Some Limits On Our Capacity For Processing Information. *Brain physiology & psychology*, *63*, 175.

Mulholland, T. M., Pellegrino, J. W. & Glaser, R. (1980). Components of geometric analogy solution. *Cognitive Psychology*, *12*(2), 252–284. doi: 10.1016/0010-0285(80)90011-0

NetApplications.com. (2018). *Browser Share by Version.* Retrieved 2018-09-13, from `www.netmarketshare.com`

Primi, R. (2002). Complexity of geometric inductive reasoning tasks contribution to the understanding of fluid intelligence. *Intelligence*, *30*(1), 41–70. doi: 10.1016/S0160-2896(01)00067-8

Puddey, I. B., Mercer, A., Andrich, D. & Styles, I. (2014). Practice effects in medical school entrance testing with the undergraduate medicine and health

sciences admission test (UMAT). *BMC Medical Education*, *14*(1), 1–15. doi: 10.1186/1472-6920-14-48

Ragni, M. & Neubert, S. (2014). Analyzing Raven's Intelligence Test: Cognitive Model, Demand, and Complexity. In H. Prade & R. Gilles (Eds.), *Computational approaches to analogical reasoning: Current trends* (pp. 351–370). Springer.

Ragni, M., Stahl, P. & Fangmeier, T. (2011). Cognitive Complexity in Matrix Reasoning Tasks. *European Perspectives on Cognitive Science*(1972).

Raven, J. (1958). *Raven's Standard Progressive Matricies*.

Raven, J. (2000). The Raven's Progressive Matrices: Change and Stability over Culture and Time. *Cognitive Psychology*, *41*(1), 1–48. doi: 10.1006/cogp.1999.0735

Raven, J., Court, J. & Raven Jr., J. (1983). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Standard Progressive Matrices, (Sect 1.)*. H.K. Lewis.

Raven, J., Prieler, J. & Benesch, M. (2005). A Replication and Extension of the Item-Analysis of the Standard Progressive Matrices Plus, Together With a Comparison of the Results of Applying Three Variants of Item Response. *WPE WebPsychEmpiricist*.

Reeve, C. L. & Lam, H. (2005). The psychometric paradox of practice effects due to retesting: Measurement invariance and stable ability estimates in the face of observed score changes. *Intelligence*, *33*(5), 535–549. doi: 10.1016/j.intell.2005.05.003

Rentz, D. M., Huh, T. J., Faust, R. R., Budson, A. E., Scinto, L. F. M., Sperling, R. a. & Daffner, K. R. (2004, jan). Use of IQ-Adjusted Norms to

Predict Progressive Cognitive Decline in Highly Intelligent Older Individuals. *Neuropsychology*, *18*(1), 38–49. doi: 10.1037/0894-4105.18.1.38

Rodriguez, M. C. (2005). Three Options Are Optimal for Multiple-Choice Items: A Meta-Analysis of 80 Years of Research. *Educational Measurement Issues and Practice*, *24*(2), 3–13.

Roediger, H. L. & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests imporves long-term retention. *Psychological Science*, *17*(3), 249–255. doi: 10.1111/j.1467-9280.2006.01693.x

Salthouse, T. A. (2012). Robust Cognitive Change. *Journal of the International Neuropsychological Society*, *18*(4), 749–756.

Salthouse, T. A. (2013). Effects of age and ability on components of cognitive change. *Intelligence*, *41*(5), 501–511. Retrieved from `http://dx.doi.org/10.1016/j.intell.2013.07.005` doi: 10.1016/j.intell.2013.07.005

Salthouse, T. A. (2015). Implications of the Flynn effect for age-cognition relations. *Intelligence*, *48*, 51–57. Retrieved from `http://dx.doi.org/10.1016/j.intell.2014.10.007` doi: 10.1016/j.intell.2014.10.007

Salthouse, T. A. (2017). Comparable Consistency, Coherence, and Commonality of Measures of Cognitive Functioning Across Adulthood. *Assessment*. doi: 10.1177/1073191117721742

Salthouse, T. A., Schroeder, D. H. & Ferrer, E. (2004). Estimating Retest Effects in Longitudinal Assessments of Cognitive Functioning in Adults Between 18 and 60 Years of Age. *Developmental psychology*, *40*(5), 813–22. doi: 10.1037/0012-1649.40.5.813

Schaie, K. W. & Hertzog, C. (1983). Fourteen-Year Cohort-Sequential Analyses of Adult Intellectual Development. *Developmental Psychology*, *19*(4), 531–543. doi: 10.1037/0012-1649.19.4.531

Scharfen, J., Peters, J. M. & Holling, H. (2018). Retest effects in cognitive ability tests: A meta-analysis. *Intelligence*, *67*(July 2017), 44–66. doi: 10.1016/j.intell.2018.01.003

Singh-Manoux, A., Ferrie, J. E., Lynch, J. W. & Marmot, M. (2005). The role of cognitive ability (intelligence) in explaining the association between socioeconomic position and health: Evidence from the Whitehall II prospective cohort study. *American Journal of Epidemiology*, *161*(9), 831–839. doi: 10.1093/aje/kwi109

Skuy, M., Gewer, A., Osrin, Y., Khunou, D., Fridjhon, P. & Rushton, J. P. (2002). Effects of mediated learning experience on Raven's matrices scores of African and non-African university students in South Africa. *Intelligence*, *30*(3), 221–232. doi: 10.1016/S0160-2896(01)00085-X

Staff, R. T., Hogan, M. J. & Whalley, L. J. (2014). Aging Trajectories of Fluid Intelligence in Late Life: The Influence of Age, Practice and Childhood IQ on Raven's Progressive Matrices. *Intelligence*, *47*, 194–201.

Stafford, T. & Dewar, M. (2014). Tracing the Trajectory of Skill Learning With a Very Large Sample of Online Game Players. *Psychological Science*, *25*(2), 511–518. doi: 10.1177/0956797613511466

Strauss, E., Sherman, E. M. S. & Otfried Spreen. (2006). *A compendium of Neuropsychological Tests: Administration, Norms and Commentary* (3rd ed.). Oxford: Oxford University Press.

Streiner, D. L., Norman, G. R. & Cairney, J. (2015). *Health Measurement Scales* (5th ed.). Oxford University Press.

te Nijenhuis, J., van Vianen, A. E. & van der Flier, H. (2007). Score gains on g-loaded tests: No g. *Intelligence*, *35*(3), 283–300.

UK Biobank. (2018). Retrieved 2018-09-21, from `http://www.ukbiobank.ac.uk/` `http://biobank.ctsu.ox.ac.uk/crystal/label.cgi`

Verguts, T. & Boeck, P. D. (2002). The induction of solution rules in Raven's Progressive Matrices Test. *European Journal of Cognitive Psychology*, *14*(4), 521–547. doi: 10.1080/09541440143000230

Villado, A. J., Randall, J. G. & Zimmer, C. U. (2016). The Effect of Method Characteristics on Retest Score Gains and Criterion-Related Validity. *Journal of Business and Psychology*, *31*(2), 233–248. doi: 10.1007/s10869-015-9408-7

Wainer, H., Dorans, N. J., Eignor, D., Flaugher, R., Green, B. F., Mislevy, R. J., ... Thissen, D. (2014). *Computerized Adaptive Testing: A Primer* (2nd ed.). Routledge.

White, A. & Zammarelli, J. (1981). Convergence Principles: Information in the Answer Sets of Some Multiple-Choice Intelligence Tests. *Applied Psychological Measurement*, *5*(1), 21–27.

Whitehall II. (2018). Retrieved 2018-09-17, from `https://www.ucl.ac.uk/iehc/research/epidemiology-public-health/research/whiteh`

Wise, S. L., Ma, L. & Theaker, R. A. (2012). Identifying NonEffortful Student Behavior on Adaptive Tests: Implications for Test Fraud Detection. *Statistical Detection of Potential Test Fraud Conference*.

Woods, D. L., Kishiyama, M. M., Yund, E. W., Herron, T. J., Edwards, B., Poliva, O., ... Reed, B. (2011). Improving digit span assessment of short-term verbal memory. *Journal of Clinical and Experimental Neuropsychology*, *33*(1), 101–111. doi: 10.1080/13803395.2010.493149

Appendix A

Studies

Example items are not included for traditional tests due to concerns about test security.

A.1 Corvus Validation Study

On arrival to the study participants were provided with an information pack (Appendix A.1.1), which included a consent form. On completion of the consent form, participants were then asked if they would like to fill out the personal details questionnaire and the Mental Health Inventory-5 (Appendix A.1.2 and Appendix A.1.3 respectively), though it was made clear that these were optional. After completion the participant proceeded with the tests (see Chapter 5, Corvus Validation).

Note that the during this study I was based at the University of Cardiff — I transferred to the University of Oxford soon afterwards, and the contact details contained in the information reflect this situation.

A.1.1 Information

The Cardiff Puzzle

What is it about?

We are interested in comparing a number of different cognitive tests. We want to use the information collected by this study to find out how good a new test is, compared with established tests. In the long run this can help to develop more effective health studies.

'The Cardiff Puzzle' is a research study run and funded by Cardiff University School of Medicine. It has been scientifically and ethically approved.

How does it work?

We will ask you to fill in a few background details: your gender, age, first language, and your education. We will then ask you about how your mood has been for the past month, as this has been shown to influence test scores, before asking you to work through the cognitive tests.

After you have finished we may ask for any feedback you might have about the any of the tests. Altogether, collecting the information will take about an hour.

You may skip answering any question – whether regarding your details, your mood or during the cognitive studies. You may also stop participating at any time and you do not need to explain your choice to do so.

If you finish the study, and answered most of the questions we will pay you £20 worth of Amazon.co.uk vouchers as thanks for your participation.

Why is it important that I take part?

The more people take part, the more reliable the study will be, as it will represent a wider range of experiences and backgrounds. We believe that this study could help improve our health research programme.

ABOUT THE TESTS

What are the tests about?

We are interested in comparing the different tests against each other. Each of them is designed to measure subsets of cognitive ability. But they go about doing so differently. If, for any of the questions you would prefer to, then you may select 'Prefer not to answer' or 'Skip this question' and move on to the next question.

How long will answering the study take?

We expect that completing all of the tests could take an about an hour without distraction.

WHO CAN TAKE PART IN THIS STUDY?

Anyone can take part provided they are aged 18 or over, can get themselves to the study, are fluent English speakers and have normal or corrected to normal vision.

RISKS AND BENEFITS

We are not aware of any risk associated with taking part in this study.

However, if you have any particular worries you can discuss them with us, and you may withdraw at any time.

CONSENT AND WITHDRAWAL

We will need your consent to be able to enter you into the study.

You will be able to withdraw from the study at any time.

If you decide to withdraw you don't have to give a reason, though it may help us if you do so, because it may be something we can change or improve in the future. This study will be more valuable with greater numbers of people completing it.

CONFIDENTIALITY

All the information you give to us will be treated as confidential.

The records of participants will be maintained in full accordance with the terms of the Data Protection Act. The security and confidentiality of your data are our top priority. We operate stringent security measures.

STUDY TEAM

John Gallacher is an epidemiologist and psychologist and a Professor at Oxford University. His research interests include psychological resources and wellbeing, psychosocial determinants of healthy ageing and the development of large-scale and web-based epidemiologic methods. He has worked for the medical research council for over 10 years and has been the principal investigator of a number of population studies including the Caerphilly Prospective Study. He is on the steering committee of UK Biobank, a gene-environment interaction study of 500,000 people and one of the most complex population studies ever undertaken.

Isaac Thimbleby is a PhD student at Cardiff University and developed the new cognitive test generator. He is a Lucent Global Science Scholar, has a degree in Mathematics and a Masters by Research in Computer Science.

MORE ABOUT THE STUDY

How is this study financed?

Isaac Thimbleby's PhD budget funds this study, including the £20 voucher for participating.

Are the tests copyrighted?

These materials are copyrighted and licensed for use only by the registered participants for this study. They may not be redistributed, decompiled or copied.

MORE ABOUT CONFIDENTIALITY

What are you going to do with the information?

The information you give us will be used for research purposes only.

How can you guarantee confidentiality?

The main methods we use to protect your confidentiality are:

(1) Personal identifiers (e.g., your name) are kept separate from the rest of the data so that researchers analysing your data will not know who you are.

(2) Access is kept to a minimum. The computers, which hold your information, are protected by industry strength firewalls, to keep them safe from hackers.

MISCELLANEOUS

My friend/colleague/relative really wants to take part as well. How can they sign up?

Please tell them to contact us at the email address, phone number or post address provided for this study.

Has this study been scientifically and ethically approved?

All clinical research is considered by an independent group of people, called a Research Ethics Committee, to protect your safety, rights, wellbeing and dignity. This study has been reviewed and given a favourable opinion by the Cardiff University Medical School Research Ethics Committee. The study was also reviewed by an independent scientist as part of the funding process.

You can find more general information about participating in health research in this PDF from the Association of Research Ethics Committees: Health Research and You.

English is not my first language. Can you send me the information in other languages?

No, unfortunately we are only able to conduct the study in English.

Do you need me because there is something wrong with me?

Absolutely not. We know nothing about your health.

You are welcome to participate as long as you comply with the eligibility criteria stated in the section 'About the study'. But if you have any particular worries about any of the aspects involved with the interventions or the questionnaires, you should discuss them with your doctor, the appropriate professional advisor or a trusted friend.

How do I make a complaint or express a concern?

Should you be concerned about any aspect of this study you can contact our office by email (thimblebyij@cardiff.ac.uk).

You can also make a complaint or express a concern in writing by letter to Dr. John Gallacher:
Dr. John Gallacher
Department of Psychiatry
Warneford Hospital
Oxford University
OX3 7JX

Or by email at john.gallacher@psych.ox.ac.uk

If you are not satisfied with the response you receive from Dr. John Gallacher, you can contact the Chair of Cardiff University Medical School Research Ethics Committee:
Dr. Andrew Freedman
Cardiff University
Heath Park
Cardiff CF14 4XN

We need your consent to be able to enter you into the study. For this, you will have to read some information in the consent form about what you will be agreeing to. Then, if you are still willing to take part, you can give us your consent by filling in your name, the date and ticking a box where indicated on the form.

CONSENT

Taking part in 'The Cardiff Puzzle' is voluntary. If you join and then change your mind, you can leave the study at any time. If it's not convenient for you to sign up right now, you can come back later if the study is still running.
If you take part in the study we will ask you some questions about your mood and then you will be asked to complete the cognitive tests in a randomly assigned order. These tests are Raven's Progressive Matrices, UK Biobank's Fluid Intelligence Test and the Matrix component of Cognito and a new matrix-based test.
We will store information about you if you provide it. We will separate names and personal details from all information so that it is anonymous. All information will be held securely by Cardiff University. The study has been scientifically and ethically approved.

By agreeing to take part you **declare** that:

- **You are fluent in English**
- **You are aged 18+**
- **You have normal or corrected to normal vision**

Now that you understand what is involved, do you agree to take part in 'The Cardiff Puzzle'?

Name: _____ Date:_____ Consent: ☐

If you don't want to join the study you don't have to do anything. Thank you for reading this far.

Contact Details

By email:

thimblebyIJ@cardiff.ac.uk

By post:

'The Cardiff Puzzle'
Institute of Primary Care and Public Care
519B, 5th floor, Neuadd Meirionnydd
Cardiff University
Heath Park
Cardiff, CF14 4YS

We will try to answer any questions you may have.

A.1.2 Personal Details

IDN:

What is your gender?

Female
Male
Other
Prefer not to say

What is your age?

What is your education level?

- ○ Up to GCSE/GCE 'O'level/CSE
- ○ Further education (e.g. 'A' levels, tertiary colleges, specialist colleges)
- ○ Higher education: Bachelor's Degree
- ○ Higher education: Postgraduate Degree
- ○ Prefer not to say

What is your first language?

A.1.3 Mental Health Inventory-5

Your mood

Please complete the following questions about your mood.

How much of the time **during the past month** have you:

Been very nervous?
- ○ All of the time
- ○ Most of the time
- ○ Some of the time
- ○ A little of the time
- ○ None of the time
- ○ Prefer not to say

Felt so down in the dumps that nothing could cheer you up?
- ○ All of the time
- ○ Most of the time
- ○ Some of the time
- ○ A little of the time
- ○ None of the time
- ○ Prefer not to say

Felt calm and peaceful inside?
- ○ All of the time
- ○ Most of the time
- ○ Some of the time
- ○ A little of the time
- ○ None of the time
- ○ Prefer not to say

Felt downhearted and depressed?
- ○ All of the time
- ○ Most of the time
- ○ Some of the time
- ○ A little of the time
- ○ None of the time
- ○ Prefer not to say

Been happy?
- ○ All of the time
- ○ Most of the time
- ○ Some of the time
- ○ A little of the time
- ○ None of the time
- ○ Prefer not to say

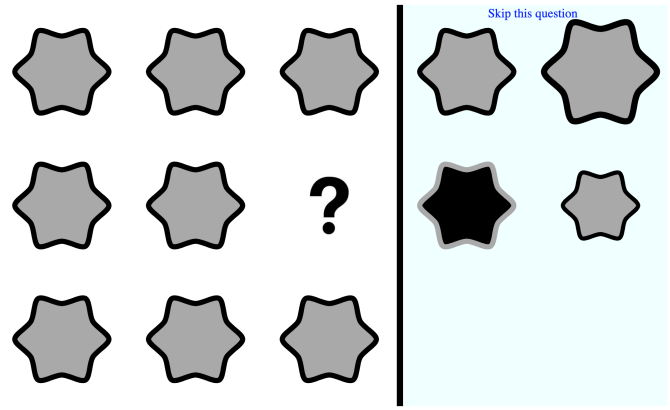
A.2 Corvus Validation Test Item Set

The test items presented below were generated using the internal version 0.6.4 of Corvus, and as such it may not be a simple task to replicate these test items exactly with a more recent version.

Each test item is presented together with the input parameters for the variable `allPuzzleTypes` that generated that test item, see section 4.1.2 for more details, as well as the comments in Corvus's code, in file `02_05-testItems.js`.

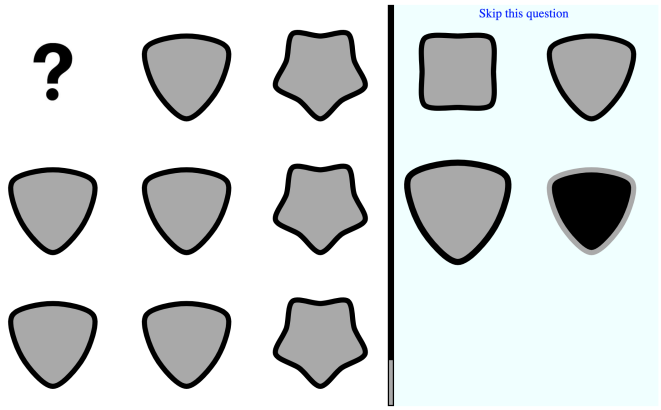
One particular way these older items differ from more recent versions of Corvus, is that the entries for Forms can not be randomised within their categories (i.e. horizontal/vertical and the two diagonals), instead they are specified specifically by an integer between one and four. As usual, zeros indicate that that feature is not used.

The items are presented here in the same order that they appeared in the test.



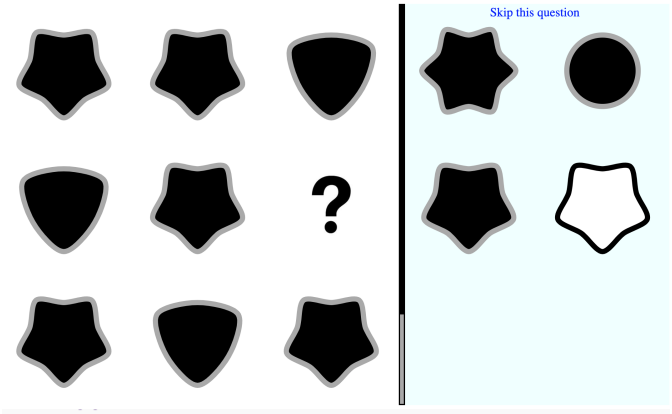

```
[[[3,3],
    [0],
    [0],
    [0,[0,0,0]],
    [[0,[1,0,0]],[1,[0,0,0]],[2,[0,0,0]]],
    [0],
    [0,0,
     0,0],
    [0,0,0],
     0,
     0],
```

Figure A.1: Most traditional Raven's-like tests start with the simplest possible item, using the identity Rule



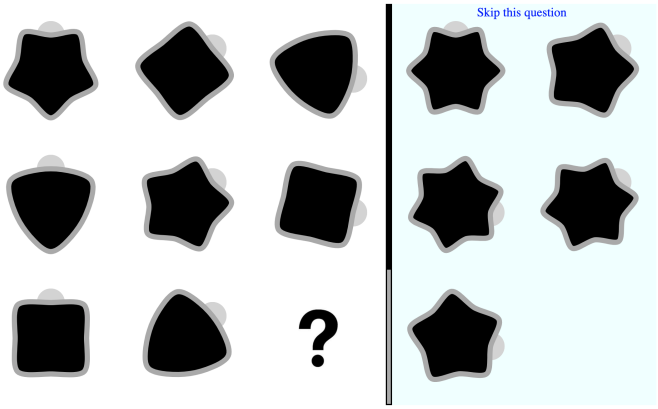

```
[[3,3],
   [0],
   [2],
   [0,[0,1,0]],
   [[0,[0,1,0]],[1,[0,0,0]],[0,[1,0,0]]],
   [8],
   [[0,1,1,1,0,1,0,0],
    [1,0,1,1,1,1,0,1],
    [1,0,1,1,1,0,0,0],
    [0,0,0,1,1,0,0,1]],
   [[1,0,0,1,1,0,0,1],
    [0,0,0,0,0,0,0,1],
    [1,0,1,1,0,1,0,0]],
    0,
    0],
```

Figure A.2: The entries in elements 5-7 (Number of elements in centre, Number Layout and Number Option Layout), are all ignored due to the Logic Setting in element 2. Normally these would be set to zero, but this item was occasionally being used to rapidly test more advanced Logic Settings



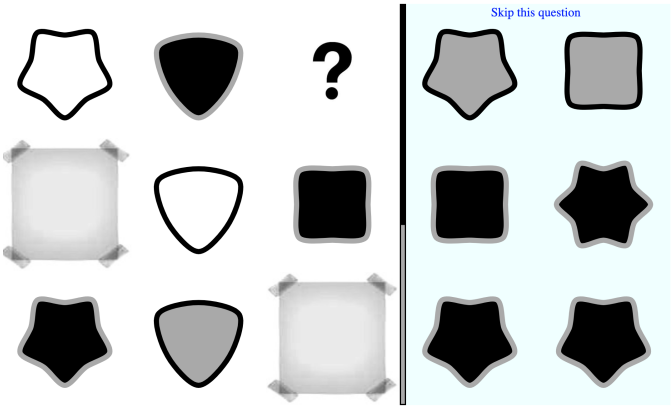

```
[[3,3],
   [0],
   [2],
   [0,[0,4,0]],
   [[0,[0,1,0]],[0,[0,2,0]],[0,[1,0,0]]],
   [0],
   [0,0,
    0,0],
   [0,0,0],
    0,
    0],
```


Figure A.3: This item uses a distribution–of–two Rule, applied in an increasing diagonal Form, using the shape Attribute. The options have three anomolies, and the correct answer is the only option that appears in the matrix



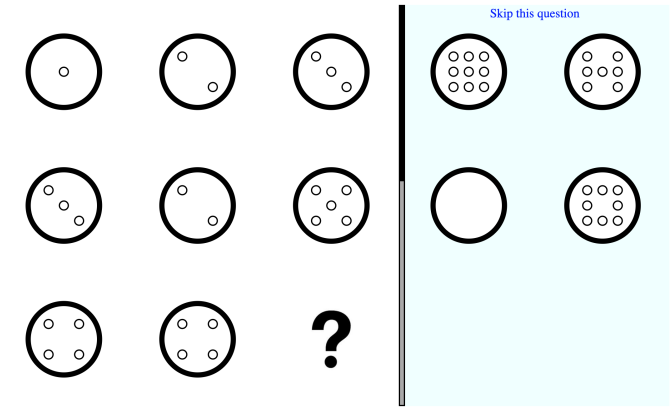

```
[[3,3],
   [0],
   [1],
   [0,[0,4,1]],
   [[0,[0,1,1]],[0,[0,1,0]],[0,[0,0,2]],[0,[0,1,2]]],
   [1],
   [3,-6,
    4,3],
   [-2,8,6],
    0,
    0],
```

Figure A.4: This test item uses two distribution–of–three Rules, where change in shape is applied in an increasing diagonal Form, and rotation is applied horizontally



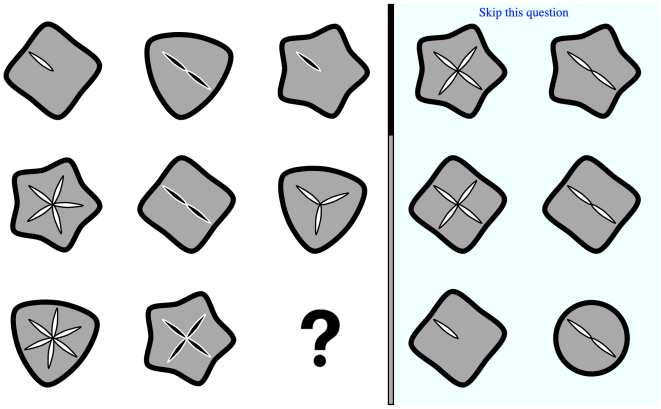

```
[[3,3],
   [0],
   [1],
   [0,[4,1,0]],
   [[0,[0,1,0]],[0,[1,0,0]],[0,[1,1,0]],[0,[1,2,0]],[0,[1,1,0]]],
   [0],
   [0,0,
    0,0],
   [0,0,0],
    2,
    0],
```

Figure A.5: This item introduces concealed elements, as per Cattell's Culture Fair



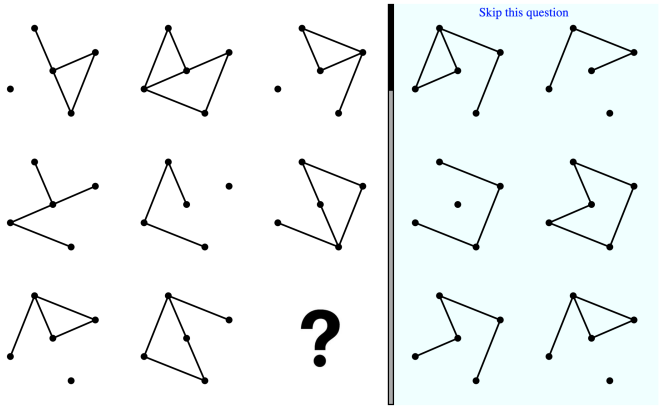

```
[[3,3],
   [1],
   [3],
   [0,[0,0,0]],
   [[0,[0,0,0]],[0,[0,0,0]],[0,[0,0,0]]],
   [1],
   [1,2,
    3,2],
   [1,2,-1],
    0,
    0],
```

Figure A.6: A test item using the Rule addition, with no negative numbers



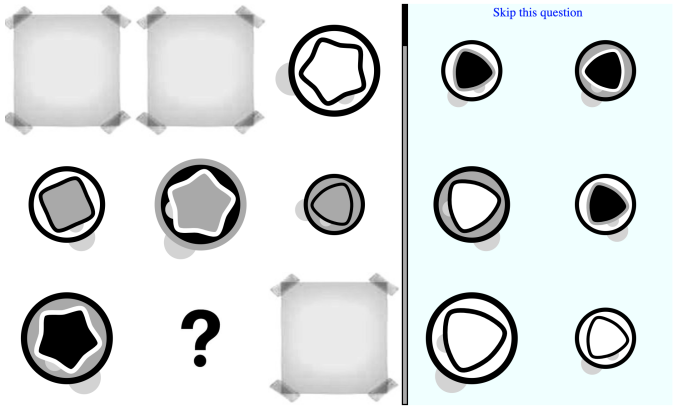

```
[[3,3],
   [2],
   [3],
   [0,[0,4,0]],
   [[0,[0,1,0]],[0,[0,0,0]],[0,[0,0,0]],[0,[0,6,0]],[0,[0,3,0]]],
   [1],
   [1,-2,
    5,-2],
   [2,2,-1,0,0],
    0,
    0],
```

Figure A.7: This item combines the previous item's Rule of addition, with both negative numbers, and a second Rule, i.e. distribution–of–three applied in an increasing diagonal using shape




```
[[3,3],
   [6],
   [6],
   [0,[0,0,0]],
   [[0,[0,0,0]],[0,[0,0,0]],[0,[0,0,0]],[0,[0,0,0]],[0,[0,0,0]]],
   [8],
   [[0,1,1,1,0,1,0,0],
    [1,0,1,1,1,1,0,1],
    [1,0,1,1,1,0,0,0],
    [0,0,0,1,1,0,0,1]],
   [[1,0,0,1,1,0,0,1],
    [0,0,0,0,0,0,0,1],
    [1,0,1,1,0,1,0,0],
    [0,0,0,0,1,0,0,1],
    [1,0,1,0,0,1,0,0]],
    0,
    0],
```

Figure A.8: The penultimate test item uses a Logic Gate, in this case NXOR. Note that there was a bug in version 0.6.4 of Corvus that caused the Logic Gate used to be one higher than it should be (according to the specification for Logic setting). i.e. a logic setting of six should mean XOR is used




```
[[3,3],
   [0],
   [1],
   [3,[4,0,1],[2,3,4]],
   [[1,[1,0,2],[1,0,2]],[2,[0,0,0],[1,0,2]],[0,[1,0,0],[0,0,1]],[0,[0,0,2],[0,0,0]],[0,[0,0,0],[1,0,0]]],
   [0],
   [0,0,
    0,0],
   [0,0,0],
    3,
    0]];
```

Figure A.9: This final item combines concealed elements with multiple distribution–of–three across two annuli. All but one distribution Rule that could be used is under use here, however as there are only four Forms available, this means that some Rules map exactly onto each other, such as the shade of the outer annuli, and the rotation of the inner annuli

A.3 True Colours

Unlike the Corvus Validation Study, The Learning Effects Study and Mouseover Study were both conducted online via an online platform called True-Colours.

True Colours is a secure platform, that enabled each participant to log in and access the questionnaires and tests I was asking them to complete. It also provided a number of other features, such as providing all participants with access to their own results, and giving additional means for the participants to contact me.

True Colours was designed to enable NHS patients to monitor their own health over time, via questionnaires and online tests, and to provide remote access to that data to the participant's clinicians.

For the purposes of this thesis True Colours handled the security and data management side of my online studies.

More information about True Colours can be found here:

`https://innovation.ox.ac.uk/licence-details/truecolours/`

A.4 Learning Effects Study

A.4.1 Information

University of Oxford *Department of Psychiatry, Warneford Hospital, University of Oxford, OX3 7JX*
Email: Isaac.thimbleby@stcatz.ox.ac.uk

Corvus Research Program: Quantifying learning effects in Cognitive Testing

What is it about?

We are interested in looking at the learning effect in repeat cognitive testing. We want to use the information collected to see if frequent testing is possible in screening for brain health.

This study is part the Corvus research program run by Isaac Thimbleby, a DPhil student at the University of Oxford, and will form part of his thesis. It has been scientifically reviewed and ethically approved.

How does it work?

First we will need you to provide consent to join the study, for which we will provide a form. You can give us your consent by ticking the boxes where indicated on the form.

The next thing to happen will be registering an account on the TrueColours website. As part of doing so, you will be asked to provide an email address and a password. Any emails you receive from the researchers will be addressed to that email.

We will then ask a few background details regarding: your gender, age, first language, and your education. We wont ask for your name here, and your name will not be included in any analysis. Next we will ask you a few questions on mood. Then you will be taken to the first cognitive test.

You may skip answering any question – whether regarding your details, your mood or during the cognitive study. You may also stop participating at any time and you do not need to explain your choice to do so.

Over the next two weeks you will be asked to log back in to complete a cognitive test for a total of 5 times. We would prefer you to do this within 24 hours of receiving the email. It is estimated that

each test will take an average of one hour, for an estimated total commitment of 5 hours over two weeks.

As a token of our gratitude for your participation, you will have the opportunity to win Amazon.co.uk vouchers to the value of £100, £40, or one of three at £20. This will be sent to you via email.

Why is it important that I take part?

Developing tools to assess brain health is important for studying cognitive decline and dementia. The tool being developed in the Corvus Program will enable changes in brain health to be detected more quickly and more accurately.

ABOUT THE TESTS

What are the tests about?

The tests are designed to assess problem solving. We are interested in how participant's answers to the tests change over time with repeat testing.

How long will answering the study take?

We expect that completing each of the tests could take an about an hour without distraction. Participants will be asked to complete a total of 5 such tests spread over two weeks.

WHO CAN TAKE PART IN THIS STUDY?

Anyone can take part provided they are aged 18 or over, have access to a computer with a mouse or trackpad and access to the Internet, English is your first language and have normal or corrected to normal vision.

Unfortunately touch screens wont work with the test.

RISKS AND BENEFITS

We are not aware of any risk associated with taking part in this study.

However, if you have any particular worries you can discuss them with us, and you may withdraw at any time.

We will delete your user account three months after the study has finished (Your username, password and email address) or sooner at your request. However this does not prevent you from contacting the researchers involved, and you may continue to do so.

Benefits include the challenge of doing the test and contributing to science.

CONSENT AND WITHDRAWAL

We will need your consent to be able to enter you into the study, entering the study is entirely optional and you are free to choose not to do so.

You will be able to withdraw from the study at any time and without penalty by advising the researchers of this decision.

If you decide to withdraw you don't have to give a reason, though it may help us if you do so, because it may be something we can change or improve in the future. This study will be more valuable with greater numbers of people completing it.

University of Oxford *Department of Psychiatry, Warneford Hospital, University of Oxford, OX3 7JX*
Email: Isaac.thimbleby@stcatz.ox.ac.uk

CONFIDENTIALITY

All the information you give to us will be treated as confidential and only Isaac Thimbleby will have access to all of the data. Anonymised data may be shared with his DPhil advisors and will be included in his DPhil thesis.

The records of participants will be maintained in full accordance with the terms of the Data Protection Act. The security and confidentiality of your data are our top priority. We operate stringent security measures.

STUDY TEAM

John Gallacher is an epidemiologist and psychologist and a Professor at the University of Oxford. His research interests include psychological resources and wellbeing, psychosocial determinants of healthy ageing and the development of large-scale and web-based epidemiologic methods. He has worked for the medical research council for over 10 years and has been the principal investigator of a number of population studies including the Caerphilly Prospective Study. He is on the steering committee of UK Biobank, a gene-environment interaction study of 500,000 people and one of the most complex population studies ever undertaken.

Isaac Thimbleby is a DPhil student at the University of Oxford and developed the new cognitive test generator. He is a Lucent Global Science Scholar, has a degree in Mathematics and a Masters by Research in Computer Science.

University of Oxford *Department of Psychiatry, Warneford Hospital, University of Oxford, OX3 7JX*
Email: Isaac.thimbleby@stcatz.ox.ac.uk

MORE ABOUT THE STUDY

How is this study financed?

Isaac Thimbleby's DPhil budget funds this study.

Are the tests copyrighted?

These materials are copyrighted and licensed for use only by the registered participants for this study. They may not be redistributed, decompiled or copied.

MORE ABOUT CONFIDENTIALITY

What are you going to do with the information?

The information you give us will be used for research purposes only.

How can you guarantee confidentiality?

The main methods we use to protect your confidentiality are:

(1) Personal identifiers (e.g., your email address) are kept separate from the rest of the data and will be deleted three months after the study has finished.

(2) Access is kept to a minimum. The computers, which hold your information, are protected by industry strength firewalls, to keep them safe from hackers.

MISCELLANEOUS

My friend/colleague/relative really wants to take part as well. How can they sign up?

Please tell them to contact us at the email address or post address provided for this study.

Has this study been scientifically and ethically approved?

All clinical research is considered by an independent group of people, called a Research Ethics Committee, to protect your safety, rights, wellbeing and dignity. This study has been reviewed by and received ethics clearance through, the University of Oxford Central University Research Ethics Committee.

English is not my first language. Can you send me the information in other languages?

No, unfortunately we are only able to conduct the study in English.

Do you need me because there is something wrong with me?

Absolutely not. We know nothing about your health.

You are welcome to participate as long as you comply with the eligibility criteria stated in the section 'About the study'.

How do I make a complaint or express a concern?

If you have a concern about any aspect of this project, please speak to Isaac Thimbleby (Isaac.thimbleby@psych.ox.ac.uk), who will do his best to answer your query. The researcher should acknowledge your concern within 10 working days and give you an indication of how he intends to deal with it.

You can also make a complaint or express a concern in writing by letter to his DPhil supervisor:
Prof. John Gallacher
Department of Psychiatry
Warneford Hospital
University of Oxford
OX3 7JX

Or by email at john.gallacher@psych.ox.ac.uk

If you remain unhappy or wish to make a formal complaint, please contact the chair of the Research Ethics Committee at the University

of Oxford who will seek to resolve the matter in a reasonably expeditious manner:

Research Services
University of Oxford
Wellington Square
Oxford
OX1 2JD

Or by email at ethics@medsci.ox.ac.uk

Contact Details

By email:

Isaac.thimbleby@stcatz.ox.ac.uk

By post:

'Learning effects in Cognitive Testing Pilot Study'
Isaac Thimbleby
Department of Psychiatry
Warneford Hospital
Oxford
OX3 7JX

We will try to answer any questions you may have.

A.4.2 Consent

CONSENT

Taking part in 'Corvus Research Program: Quantifying learning effects in Cognitive Testing' is voluntary. If you join and then change your mind, you can leave the study at any time. If it's not convenient for you to sign up right now, you can come back later if the study is still running.

Taking part involves answering some questions about your mood and then completing a problem-solving test 5 times over the next two weeks.

We will store the personal information you provided for the duration of the study, and we will erase this information within three months of the study finishing. We will separate personal details from all cognitive data so that test performance is de-identified. All information will be held securely by the University of Oxford. The study has been scientifically and ethically approved.

By agreeing to take part you confirm, by checking the boxes provided, that:

- **You have read and understood the participant information page**
- **That you understand how to raise a concern and make a complaint**
- **That you have had the opportunity to ask questions and have received satisfactory answers to your questions**

To confirm the statements above, please check this box: ☐

- **That you may withdraw from the study without penalty at any time by advising the researchers of this decision**
- **You understand that this project has been reviewed, and received ethics clearance through the University of Oxford Central University Research Ethics Committee**
- **That you understand who will have access to any data personal data, how that data will be stored; and what will happen to the data at the end of this study**

- **That you understand that anonymised data from this study will be included in Isaac Thimbleby's DPhil student thesis**

To confirm the statements above, please check this box: ☐

- **That English is your first language**
- **You are aged 18+**
- **You have normal or corrected to normal vision**
- **You will use a mouse or track-pad to complete the test**

To confirm the statements above, please check this box: ☐

If you don't want to join the study you don't have to do anything. Thank you for reading this far.

If you do want to give confirm the statements above, please check the boxes above, save the document and return it to Isaac.Thimbleby@psych.ox.ac.uk

A.5 Mouseover Study

A.5.1 Information

University of Oxford *Department of Psychiatry, Warneford Hospital, University of Oxford, OX3 7JX*
Email: Isaac.thimbleby@psych.ox.ac.uk

Corvus Research Program: Corvus Mouseover Study
What is it about?

We are interested in looking at the effect of working memory on cognitive testing. We want to use the information collected to improve our tests, and to assist in analysing the difference between tests used in other studies.

This study is part the Corvus research program run by Isaac Thimbleby, a DPhil student at the University of Oxford, and will form part of his thesis.

WHO CAN TAKE PART IN THIS STUDY?

Participation in this study is for adult participants who are 18+, have English as their first language, normal or corrected to normal vision, access to the Internet, and the ability to use a mouse or track-pad to complete the test.

It is important to note that touch screens wont work with the test, due to the use of a mouseover feature, and that participants are asked to complete the test using a mouse or track-pad.

How does it work?

First we will need you to provide consent to join the study, for which we will provide a form. You can give us your consent by ticking the boxes where indicated on the form and sending it via a social media private message, or other encrypted messaging service to Isaac Thimbleby.

You will then be signed up to the website 'https://oscar.psych.ox.ac.uk/corvus/en/', and provided with specific instructions for accessing it via email. The tests used in this study can be found after logging in, under the 'new' tab.

You will first be asked to complete some optional personal details, which ask you for your age, sex, highest education level, as well as five mental-health questions relating to the last month. These details are recorded firstly to make sure that we have a wide range of participants, and secondly because mental-health has been shown to influence cognitive test results, such as those used in the main part of this study.

University of Oxford *Department of Psychiatry, Warneford Hospital, University of Oxford, OX3 7JX*
Email: Isaac.thimbleby@psych.ox.ac.uk

After that is complete we would like you to complete the test subtitled 'Working Memory Test', followed by the test subtitled 'Corvus Study Test'.

You may skip answering any question – regardless of it being during the personal details section, the working memory test or the Corvus tests. You may also stop participating at any time and you do not need to explain your choice to do so.

Once the study is completed all participants will be sent a £5 Amazon.co.uk voucher as compensation for their time.

Why is it important that I take part?

Developing tools to assess brain health is important for studying cognitive decline and dementia. The tool being developed in the Corvus Program will enable changes in brain health to be detected more quickly and more accurately.

ABOUT THE TESTS

What are the tests about?

The tests are designed to assess problem solving and working memory. We are interested in how participant's answers to the tests are influenced by their working memory and the mouseover feature.

How long will answering the study take?

We expect that completing both tests could take you less than an hour without distraction.

RISKS AND BENEFITS

We are not aware of any risk associated with taking part in this study. However, if you have any particular worries you can discuss them with us, and you may withdraw at any time.

We will delete your user account three months after the study has finished (Your username, password and email address) or sooner at your request. However this does not prevent you from contacting the researchers involved, and you may continue to do so.

Benefits include the challenge of doing the test and contributing to science.

University of Oxford *Department of Psychiatry, Warneford Hospital, University of Oxford, OX3 7JX*
Email: Isaac.thimbleby@psych.ox.ac.uk

CONSENT AND WITHDRAWAL

We will need your consent to be able to enter you into the study; entering the study is entirely optional and you are free to choose not to do so.

You will be able to withdraw from the study at any time and without penalty by advising the researchers of this decision.

If you decide to withdraw you don't have to give a reason, though it may help us if you do so, because it may be something we can change or improve in the future. This study will be more valuable with greater numbers of people completing it.

CONFIDENTIALITY

All the information you give to us will be treated as confidential and only Isaac Thimbleby will have access to all of the data. Anonymised data may be shared with his DPhil advisors and will be included in his DPhil thesis.

The records of participants will be maintained in full accordance with the terms of the Data Protection Act. The security and confidentiality of your data are our top priority. We operate stringent security measures.

STUDY TEAM

John Gallacher is an epidemiologist and psychologist and a Professor at the University of Oxford. His research interests include psychological resources and wellbeing, psychosocial determinants of healthy ageing and the development of large-scale and web-based epidemiologic methods. He has worked for the medical research council for over 10 years and has been the principal investigator of a number of population studies including the Caerphilly Prospective Study. He is on the steering committee of UK Biobank, a gene-environment interaction study of 500,000 people and one of the most complex population studies ever undertaken.

Isaac Thimbleby is a DPhil student at the University of Oxford and developed the new cognitive test generator. He is a Lucent Global Science Scholar, has a degree in Mathematics and a Masters by Research in Computer Science.

MORE ABOUT THE STUDY

How is this study financed?

Isaac Thimbleby's DPhil budget funds this study.

Are the tests copyrighted?

These materials are copyrighted and licensed for use only by the registered participants for this study. They may not be redistributed, decompiled or copied.

MORE ABOUT CONFIDENTIALITY

What are you going to do with the information?

The information you give us will be used for research purposes only.

How can you guarantee confidentiality?

The main methods we use to protect your confidentiality are:

(1) Personal identifiers (e.g., your email address) are kept separate from the rest of the data and will be deleted three months after the study has finished.

(2) Access is kept to a minimum. The computers, which hold your information, are protected by industry strength firewalls, to keep them safe from hackers.

MISCELLANEOUS

Has this study been scientifically and ethically approved?

All clinical research is considered by an independent group of people, called a Research Ethics Committee, to protect your safety, rights, wellbeing and dignity. This study has been reviewed by and received ethics clearance through, the University of Oxford Central University Research Ethics Committee.

English is not my first language. Can you send me the information in other languages?

No, unfortunately we are only able to conduct the study in English.

Do you need me because there is something wrong with me?

Absolutely not. We know nothing about your health.

You are welcome to participate as long as you comply with the eligibility criteria stated in the section 'About the study'.

How do I make a complaint or express a concern?

If you have a concern about any aspect of this project, please speak to Isaac Thimbleby (Isaac.thimbleby@psych.ox.ac.uk), who will do his best to answer your query. The researcher should acknowledge your concern within 10 working days and give you an indication of how he intends to deal with it.

You can also make a complaint or express a concern in writing by letter to his DPhil supervisor:
Prof. John Gallacher
Department of Psychiatry
Warneford Hospital
University of Oxford
OX3 7JX

Or by email at john.gallacher@psych.ox.ac.uk

If you remain unhappy or wish to make a formal complaint, please contact the chair of the Research Ethics Committee at the University of Oxford who will seek to resolve the matter in a reasonably expeditious manner:
Research Services
University of Oxford
Wellington Square
Oxford
OX1 2JD

Or by email at ethics@medsci.ox.ac.uk

Contact Details

By email:

Isaac.thimbleby@psych.ox.ac.uk

By post:

'Third Corvus Study'
Isaac Thimbleby
Department of Psychiatry
Warneford Hospital
Oxford
OX3 7JX

We will try to answer any questions you may have.

A.5.2 Consent

Department of Psychiatry

John Gallacher
John.Gallacher@psych.ox.ac.uk
Isaac Thimbleby, DPhil Student
Oxford telephone number: 01865 xxxxxx
Oxford e-mail: isaac.thimbleby@psych.ox.ac.uk

PARTICIPANT CONSENT FORM

CUREC Approval Reference: R54659/RE001

Corvus Research Program: Corvus Mouseover Study

Purpose of Study: Investigating the interaction between Corvus and working memory.

		Please initial each box
1	I confirm that I have read and understand the information for the above study. I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily.	☐
2	I understand that my participation is voluntary and that I am free to withdraw at any time, without giving any reason, and without any adverse consequences or academic penalty.	☐
3	I understand that research data collected during the study may be looked at by designated individuals from the University of Oxford where it is relevant to my taking part in this study. I give permission for these individuals to access my data.	☐
4	I understand that this project has been reviewed by, and received ethics clearance through, the University of Oxford Central University Research Ethics Committee.	☐
5	I understand who will have access to personal data provided, how the data will be stored and what will happen to the data at the end of the project.	☐
6	I understand how this research will be written up and published.	☐
7	I understand how to raise a concern or make a complaint.	☐
8	I am 18+ years old	☐
9	I have access to and use of a computer and the internet	☐
10	I have normal or corrected to normal vision	☐
11	English is my first language	☐
12	I agree to complete the study using either a mouse or a track pad	☐

13	I agree to take part in the study	☐

The following statements are optional

14	I agree for research data collected in this study to be given to researchers, including those working outside of the EU, to be used in other research studies. I understand that any data that leave the research group will be fully anonymised so that I cannot be identified.	☐
15	I agree for my personal data to be kept in a secure database for the purpose of contacting me about future studies.	☐

_____	dd / mm / yyyy	_____
Name of Participant	Date	Signature

_____

Email address of Participant

_____	dd / mm / yyyy	_____
Name of person taking consent	Date	Signature

Appendix B

Code

All code that has been made generally available has been uploaded to:

`https://github.com/Thimbleby?tab=repositories`

Please see the files themselves for the licences under which they have been made available. Some files include publicly available libraries which may come with their own open-source licensing. Additional code, such as the exact version of Corvus used in a particular study, can be made available on request.

Note that deploying the version of Corvus uploaded on of 28th of September 2018 requires someone confident in programming in JavaScript. The version uploaded at that time is the version described in Chapter 7, Mouseover Study, and tailored for use with TrueColours.

It is also important to note that the focus in designing Corvus, to date, had not been to design the perfect Raven's-like test generator, but rather to design one that could emulate Raven's SPM and others to a controllable degree. It is likely that any future uploads of Corvus will widen the scope of this goal, as it is no longer constrained by the context of this thesis.

Additionally some terminology in the code, particularly in the comments (as uploaded at that date), is archaic from the perspective of this thesis. Such as using size for magnitude, layout for form, and answers for options. This change in terminology occurred over the four years taken to complete this DPhil.

A user manual has been provided, and can be, and can also be found at the same github repository.

B.1 User Manual

Corvus User Manual

Corvus is a Raven's-like test generator written in JavaScript. Further details on it, including an explanation of some of the terms used here, can be found in the author's DPhil "Corvus: an Automatic Raven's-like Test Generator", specifically chapters 1 and 4.

Corvus demonstrates the development of an approach to performing cognitive experiments at scale. Corvus can be downloaded from https://github.com/Thimbleby/Corvus. It can be launched in a browser by opening index.html, which will provide basic instructions, and a link to start the test.

As indicated by the version number, Corvus is in ongoing development. There are multiple ways in which Corvus could be improved or altered to fit specific tasks. Corvus is an experimental system designed to enable the studies mentioned in the thesis, and for many other purposes.

It is assumed that investigators will wish to tailor Corvus to their own purpose and need. This manual is written as a brief introduction to help assist users in that endeavour, however this manual does assume that such investigators are comfortable reading and editing JavaScript as a minimum requirement.

At time of writing this user manual (2019), the version of Corvus currently provided on GitHub (v0.8.8) will work on local machines and will download the results once the test is completed to the user's device. This download facility works on Google Chrome. However, while there are known issues with downloading the results on some other browsers, this is not considered an issue as this set-up was designed for testing purposes. It is presumed that a user will want to alter Corvus to record participant data to their own secure database, and this is what has been done for the author's studies.

The code that triggers downloading the test results can be found in '…/JavaScript/08-main.js' by searching for '// ### DOWNLOAD FUNCTION ###'. This line calls code that can be found in "…/Other files/download.js" — once you have implemented your own backend solution, "download.js" can be removed.

For further guidance on specifically tailoring Corvus to a user's own needs, user's may please feel free to contact the author by opening an issue on the GitHub repository for Corvus.

General Settings
Some general settings can be altered in "JavaScript/01-properties.js". Specifically, the variable 'currentSet' (Search for '// ### Mouse-over ###'), defines how the options interact with mouseovers. See the comments on preceding lines, or chapter 7 of the author's thesis for more details. This file also defines general properties such as the size of icons and the thickness of lines.

Tailoring Test Items
The key files to be edited start with the prefix '02_', and possibly parts of the file '04-pattern.js'. Editing the graphical appearance of items can be done via files with the prefix '06_'.

The file '02_04-testItems.js' handles definitions for fixed items. Elements of these fixed items can then overwritten by the functions in files '02_02-RNGAnulus.js', '02_03-RNGAdd.js', '02_04-RNGLG.js'. The file '02_01-orders.js' is primarily a commented-out function that can be used to calculate an array that generally does not change test-to-test (the results of which have been hard-coded at the end of this file as a result), as well as a function for shuffling arrays. The rest of this section of the manual describes '02_05-testItems.js'.

The file '02_05-testItems.js' begins with a large comment detailing some of the options available for each test item. Some of the options listed were included as future proofing for potential ways to expand on Corvus. Specifically, sizes of Grid other than [3,3], options 8-22 under Logic Options, and option 1 under type are not implemented as of Corvus version 0.8.8.

Note that in the code, 112 is shorthand for distribution of 2, and 123 is shorthand for distribution of 3. Similarly, 'Add' is sometimes used as shorthand for Addition.

This is followed by the array AllPuzzleTypes. This array is where each test item is specified. Further details on its structure, beyond those presented here, can also be found in section 4.2 of "Corvus: An Automatic Raven's-like Test Generator".

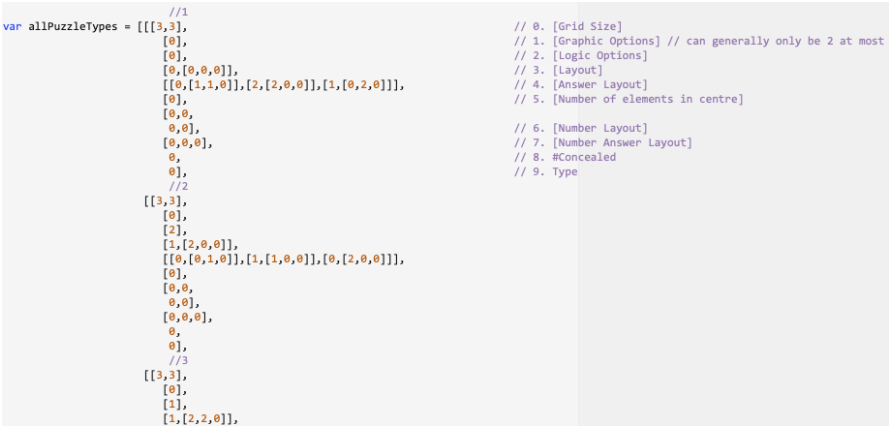


Figure 1: A screenshot of the first few elements of allPuzzleTypes.

This array is followed by a for loop, which iterates through each test item, and replaces parts of each test item with output from the functions in the other JavaScript files starting with "02".

If each participant is to take the exact same test, the random number generator seed can be fixed in "…/random.html" (Search for '// ### Seed ###'), to Math.seedrandom("X") where "X" is any fixed string. If participants are to take very similar tests this loop can instead be disabled (Search for '// ### Override allPuzzleTypes ###'), and the test items to be taken are specified in AllPuzzleTypes. By combining both of these modifications it is ensured that all participants take the same test items as specified by AllPuzzleTypes.

The reason both modifications are necessary is that even without that for loop AllPuzzleTypes does not fully specify all items. Some parts of each test item are left to random number generators, such as if a test item uses squares, triangles or circles. Similarly, the order options are presented in is randomised.

By default Layout, Answer Layout, Number of elements in Centre, Number Layout, and Number Answer Layout are all overwritten by functions after allPuzzleTypes has been initially defined. If there is a need to specify the exact test participants are asked manually, commenting out much of the section immediately following allPuzzleTypes should work.

Graphic Options
The length of this array defines the number of patterns types used. Corvus 0.8.8 is currently only intended to handle one or two pattern types, and if two are needed, one of them – and only one – must be Annulus (i.e. 0).

The *graphic option* chosen must be able to support the *logic option* chosen next. The following table indicates compatibility between *graphic options* and *logic options*. A question mark indicates that that combination of parameters has not been tested with the current version of Corvus, but that it should work in principle or with a small number of alterations or updates to the code.

	Identity	Distributions	Addition	Logic Gates
Annulus	✓	✓		
Dice	?	?	✓	
Petals	?	?	✓	
Spike Rings	?	?	?	
Tessellating Squares	?	?	?	
Tessellating Triangles	?	?	?	
BoxLines	?	?	?	✓

As may be clear from the number of question marks in the table above, this was not considered a priority for the current version, as a minimum of only one *graphic option* is needed for each *logic option*.

These *graphic options* are defined in the JavaScript files starting with the prefix "06".

Logic Options
This array should be the same length as the *Graphical Options* array, with each element of each array corresponding sequentially to each other.

There are seven logic options currently available. The *logic option* chosen here defines if and how the following six options are used, as per the table below.

	Identity	112	123	Addition	AND	OR	XOR	XNOR
Layout		✓	✓					
Answer Layout		✓	✓					
Number of elements in centre				~	✓	✓	✓	✓
Number layout				✓	✓	✓	✓	✓
Number answer layout				✓	✓	✓	✓	✓
#Concealed		✓	✓					

At present, parameters with ticks in the same row are mutually exclusive. This table — and how it combines with the previous table — is why *Graphic Options* with two pattern types must incorporate one using Annuli, and the other not.

An example combining 123 and XOR can be found towards the end of this manual.

Layout
Layout is an array in the form [magnitude, [colour, shape, rotation]] or [magnitude, [colour, shape, rotation], [colour, shape, rotation]], depending on the number of annuli wanted.

Note that Corvus does not check coherence. If the same layout for two different rules is used, it will give the test item solving redundancy; i.e. multiple independent ways of solving the same test item.

Each entry in the array defines the form the rule defined by *Logic Options*. A zero entry defines an Identity – or arguably, a lack of pattern. An entry of 1 or 2 randomly defines an orthogonal (horizontal or vertical) form; if both 1 and 2 then Corvus ensures that they are different to each other, but still randomly assigned to vertical or horizontal. Similarly, if two entries are the same, Corvus ensures that they are assigned to the same resulting form. Entries of 3 or 4 work similarly, but for increasing and decreasing diagonal forms.

If two pattern types are to be combined, it is advised that two annuli are not to be used (i.e. define Layout to be in the form [magnitude, [colour, shape, rotation], [colour, shape, rotation]]). While this will work, it will result in the inner most shapes being very small, which could present issues.

N.B. Layout is the word I initially used to refer to Form. Form is the test item property defined by the variable named layout.

Answer Layout
Answer Layout is an array of elements, each individually taking the same form as the Layout. The length of the array determines the number of incorrect options presented.

If Layout uses two annuli, then all elements of Answer Layout should also do the same.

However, here each non-zero number represents a delta from the correct answer. In other words, [0, [0,0,0]] would be the correct answer (for a test item using one annuli, and no other pattern type) – however the correct answers should not be included here; they will be added automatically at a later date, then the order of the answers will be shuffled.

An answer layout of [[0, [1,0,0]], [0, [0,1,0]], [0, [0,1,1]]] would generate a test item with four options in a random order. One of which would be the correct answer. One of the alternative options would differ in a randomly determined way in regard to colour, another would differ in regard to shape, and the last would differ in the same way as the second in regard to shape, but also in rotation.

Similarly to the way entries worked in the previous parameter, non-zero digits are randomly assigned to a particular attribute value. The digit 1 assigned to colour in one alternative option, will be assigned to the same attribute value as a 1 assigned to colour in another alternative option.

NOTE: If any matrix element or option has a rotation relative to the others, for any reason, then all matrix elements and options will have a 'rotation tab' which removes rotational symmetry.

Number of Elements in Centre

This variable is used by Corvus to interpret the following two variables. It defines the number of sub-elements used in each element.

Number of elements in centre is always 1 for Addition.

It may be the case that it is possible to remove this from the code, as in, in theory, it could be made redundant. However doing so has not been a priority.

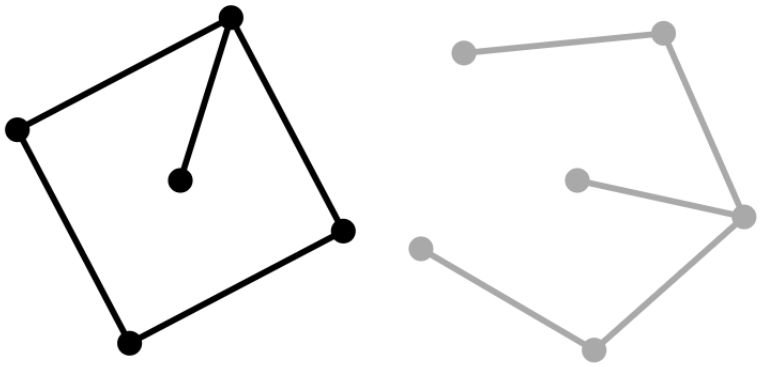


Figure 2: To generate elements like the example on the left the 'Number of Elements in Centre' should be set 5, while for generating elements like the one on the right, the same parameter should be set to 6.

Number Layout
Number Layout works slightly differently for Addition and Logic Gates. However in both cases Number Layout has 4 parts – each representing the four elements in the grid that are not in the same row or column as the missing element. If the missing element is in the bottom right-hand corner, then these arrays map onto the four elements in the top-left of the grid or matrix as follows; [top-left, top-centre, centre-left, centre].

These parts are arrays with length equal to *Number of Elements in Centre*, or integers when *Number of Elements in Centre* is one (i.e. for Addition).

With Addition, each integer is graphically encoded directly using the *Graphic Option* selected above. It is important to ensure that each of the integers, the sum of each row and column and the sum of all four integers all lie within the range that the *Graphic Option* can handle. For the Dice and Petal encodings, this range is -9 to +9. Though the Petal encoding does not have as hard a limit as the Dice encoding, that is, it could technically manage a larger range, this would not be ideal.

With Logic Gates each sub-element links to the presence or absence of a part of the graphical encoding. Zero indicates no difference between this part of the grid, and the correct answer or missing element. It is important to ensure that every permutation of Logic Gate inputs is included in the matrix, so that participants can identify, with surety, which Logic Gate is in use. This means when considering each row and column of *Number Layout*, that all four binary permutations (11, 10, 01, and 00) should all be included. Ensuring that they appear in the *Number Layout*, ensures that their output is also included; a binary permutation

appearing in the Matrix or grid, but not the Number Layout is a potential problem.

While it is possible to construct test items where this is not essential, doing so requires extra work; i.e. ensuring that every binary permutation that does not appear in *Number Layout*, also does not appear anywhere in the Matrix.

Number Answer Layout

Each element of this parameter is a distinct alternative option, and much like *Answer Layout*, the correct answer would be indicated by zeros, but is not included in this parameter as it is shuffled in at a later point.

As with *Number Layout*, this alters slightly depending on whether Addition or Logic Gates are being used. With Addition, this will be an array of integers, while for Logic Gates they will be arrays of size equal to *Number of Elements in Centre*.

For addition, each number is [double check if includes correct answer – it would be unusual for it to do so, but also I am not sure how setting 0 as the correct answer would work for addition] that element's delta from the correct answer modulated so that the end result lies in the range -9 to +9.

For Logic Gates, each 1 indicates that the sub-element indicated by its array index has the opposite presence or absence to the correct answer.

#Concealed
The concealed parameter functions with the *logic option*s for distribution ("112" and "123").

Concealed values grater than zero cause the function requiredVisSet(), in "…/JavaScript/04-pattern.js" (Search for '// ### Calculate Minimum Unconcealed ###') to run. This function attempts to conceal elements within the matrix until it either cannot conceal additional elements without compromising the test item, or until it has reached the number of elements concealed as set in this parameter.

Functions
In the file '02_05-testItem.js', after the array allPuzzleTypes is defined, a series of functions can be applied to the array — and are switched on by default.

These functions are defined in the files '02_02-RNGAnulus.js', '02_03-RNGAdd.js', and '02_04-RNGLG.js'. The files are analogous to each other in purpose, respectively for Distributions, Addition, and Logic Gates.

Each file primarily consists of two functions. The first (with the suffix 'Qu') redefines the Rule's layout.

For distributions this is very simply done using taking the definition template at position Math.floor(dif/maxDif^2) of the list of possible definitions in the hard-coded array annulusRuleArray, which is defined in '02_01-orders.js' (a lot of the commented-out code in that file can be used to calculate annulusRuleArray). Where dif/maxDif^2 is the test item number divided by the square of the total number of test items in allPuzzleTypes. If upgrading Corvus to work with Adaptive Testing, then variables found in the code such as dif and difficulty would make a good starting place for theta. The array annulusRuleArray only has values 1 for orthogonal, and 2 for diagonal. The function returns values that alternate between the two options within each category of form as it reads the generated Form from left to right, from a random starting Form subcategory.

For Addition and Logic gates, these functions generate a 2x2 array, while ensuring that every element in the full 3x3 matrix can be processed by the graphical encoding chosen. For Logic Gates, the function ensures that the full set of Boolean combinations and their results are present in the matrix, as these are necessary to fully define each Logic Gate distinctly from all other Logic Gates.

This template then replaces the original value, and similar process to that discussed under *Layout* and *Number Layout* in the previous section is applied.

The second function (with the suffix Ans) redefines a list of alternative options, i.e. *Answer Layout* and *Number Answer Layout*.

This function works by taking the number of answers determined in allPuzzleTypes, and the number of rules used.

For distributions, the function uses the number of rules to define minimum number of options that do not relate to the rules (anomalies), generating an even spread of alternative answers, generating the anomalies and then inserting them in to the array. This is necessary as with fewer rules a larger number of anomalies becomes essential. It also uses anomalies to ensure that an even distribution of values is used, in order to avoid clues and anti-clues.

For addition and logic gates the functions calculate a set distribution of answers around the correct answer, with some systematic randomisation. For Logic Gates the function also factors in test item difficulty.

Examples

Both of the test items in Figure 3, and Figure 4 were generated with the same code, which is shown in Figure 5.

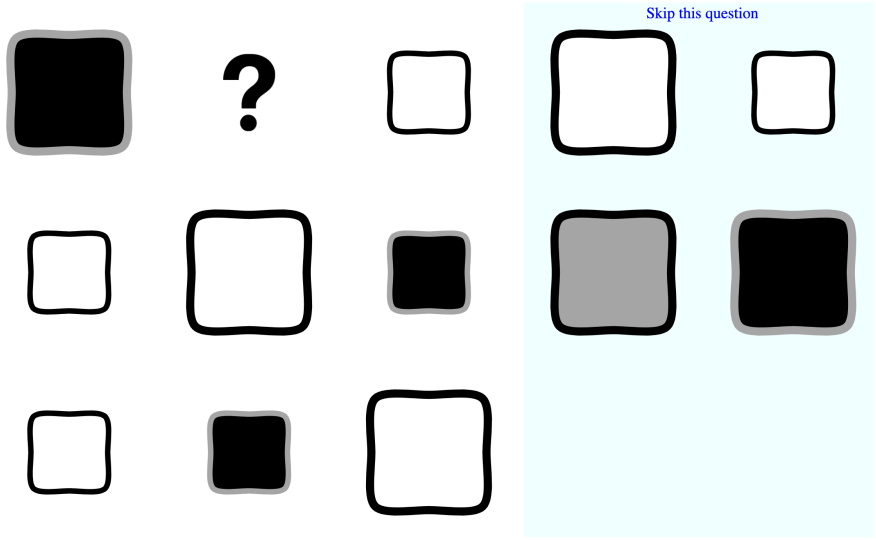


Figure 3: A screen shot of a test item with two distributions of 2.

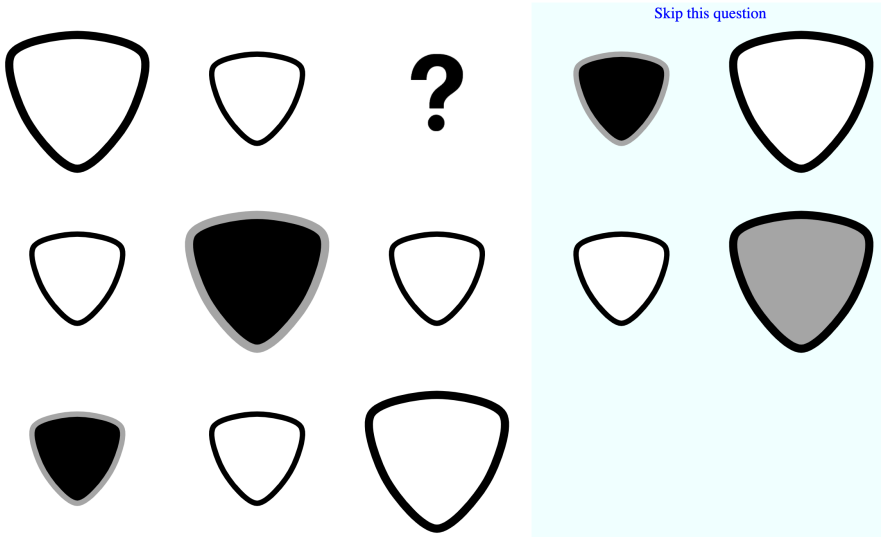


Figure 4: A screen shot with a different version of the same item as seen in Figure 1.

```
[[3,3],
   [0],
   [2],
   [1,[2,0,0]],
   [[0,[0,1,0]],[1,[1,0,0]],[0,[2,0,0]]],
   [0],
   [0,0,
    0,0],
   [0,0,0],
    0,
    0],
```

Figure 5: The AllPuzzleTypes element used to generate Figures 2 and 3.

If the functions had been turned off, this test item would have produced two distributions of 3 one working with size and the other with colour, randomly one of which would be horizontal and the other one vertical.

However the functions were active when generating the test items in Figures 2 and 3, but as the *Logic Option* was set to 2, the functions overwrite the data relevant to a Distribution of Three.

This test item was the second of ten test items in this test, so the layout or Form taken from annulusRuleArray is [2,[2,0,0]], rather than the value in allPuzzleTypes – as it happens, the only changes relate to the Forms. In annulusRuleArray, the value 2 indicates a diagonal Form, and as there are two of them, they are randomly assigned to different diagonal Forms.

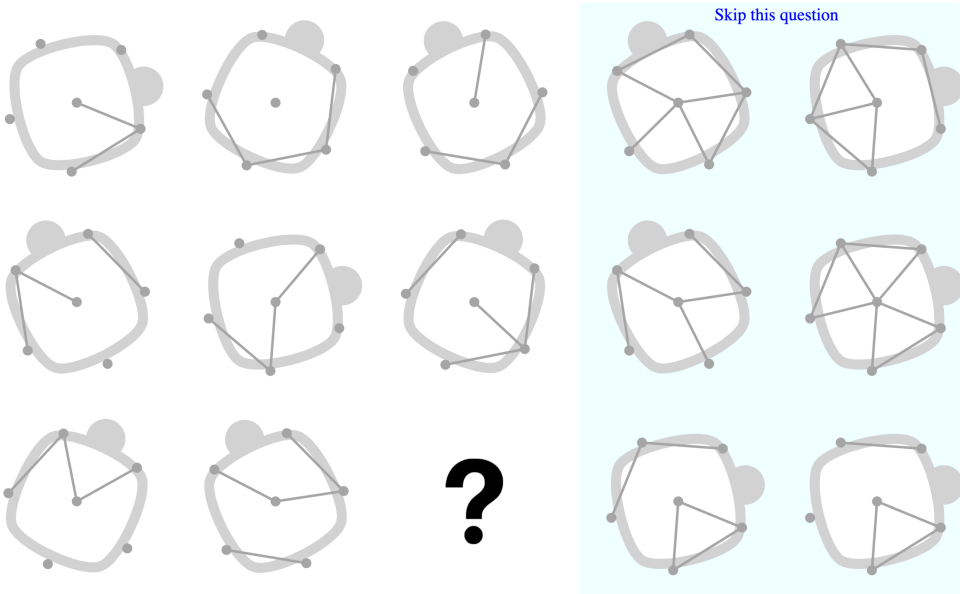


Figure 6: A screen shot of a test item combining the XOR Logic Gate with a Distribution of Three acting on rotation in an increasing diagonal Form.

The example shown in Figure 6 was generated using the code shown in Figure 7.

```
[[3,3],
   [6],
   [6],
   [0,[0,0,4]],
   [[0,[0,0,1]],[0,[0,0,1]],[0,[0,0,0]],[0,[0,0,0]],[0,[0,0,0]]],
   [10],
   [[0,1,0,0,1,1,0,1,1,0],
    [0,0,0,1,0,1,0,1,0,0],
    [1,1,0,0,0,1,0,1,0,0],
    [0,1,0,1,1,1,0,0,0,1]],
   [[0,0,0,0,0,1,0,0,0,0],
    [1,0,1,0,0,0,0,0,0,0],
    [0,0,0,1,0,0,0,0,1,1],
    [1,0,1,0,0,0,0,1,0,0],
    [1,0,0,0,0,0,0,0,1,0]],
    0,
    0],
```

Figure 7: The allPuzzleTypes element used to define the test item in Figure 6.

Figure 6 shows one of the most complex test items used in the author's thesis.

Because the L*ogic Option* is set to the XOR Logic Gate, the functions only overwrite the Logic Gate portion of the array (*Number of Elements in Centre, Number Layout, and Number Answer Layout*); the Distribution of Three rules, form, and the relevant attributes of the option set are left as they are defined in Figure 7.

It can be seen from the number of rows in *Number Answer Layout*, that there will be 5 alternative options; together with the correct answer this gives the six options we see in Figure 6. Had the functions been turned off it would have been necessary to check that every Boolean set needed to define a Logic Gate, was incorporated into the visible elements of the matrix, as defined by *Number Layout*. As it is, the functions overwrite this portion of the array, and ensure that this is maintained.

Final Word

This manual is only a very brief introduction to editing a large application. If you would like to do anything this manual does not cover, or have any questions, please feel free to contact the author by opening an issue on the GitHub repository for Corvus https://github.com/Thimbleby/Corvus.