A Novel Method for Analyzing Sequential Eye Movements Reveals the Relationship Between Learning and Strategy on Raven’s Advanced Progressive Matrices

Thesis

Presented in Partial Fulfillment of the Requirements for the Degree Master of Arts in the Graduate School of the Ohio State University

By
Taylor Ray Hayes
Graduate Program in Psychology

The Ohio State University
2011

Thesis Committee:
Alexander A. Petrov, Advisor
Per B. Sederberg
James T. Todd
© Copyright by
Taylor Ray Hayes
2011
Abstract

Eye-movement patterns contain important information about strategic information processing. Using the successor representation to capture statistical regularities in temporally extended fixation sequences it was possible to assess strategic shifts in eye-movement patterns and predict scores on Ravens Advanced Progressive Matrices (APM) test. Thirty-five participants completed two subsets of APM items on two separate days. Principal component analysis of the SRs revealed individual differences in scanning patterns. The strongest principal component quantified the tendency to scan the Raven matrix systematically; another component quantified the tendency to toggle to and from the response area. These two components predicted 56% of the variance in Raven scores. The difference in SRs also suggested that the learning effect on Raven may be due to increases in systematicity. Thus, the systematicity of eye movements is an important new strategic index on Raven that can be revealed by successor-representation analysis.
To my Mom and Dad
Acknowledgments

I would like to thank my advisor, Alexander Petrov, for all his thoughtful discussions, solid advice, and support on this project. I would also like to thank Per Sederberg for his help and insight into the successor representation, which was an integral component of the eye movement analysis.
Vita

2002 ............................... Newark High School

2006 ............................... B.A. Philosophy,
The Ohio State University

2008 to present ..................... Graduate Teaching Associate,
Department of Psychology,
The Ohio State University

Publications

Hayes, T. R. & Petrov, A. A. (2009). Perceptual learning transfers from luminance-
to contrast-defined motion. *Journal of Vision, 9*, 884.

between Luminance- and Contrast-Defined motion: Evidence for shared and distinct
processing. *Annual meeting of the Psychonomic Society*, 5095.

Petrov, A. A. & Hayes, T. R. (2010). Asymmetric transfer of perceptual learning of
Fields of Study

Major Field: Psychology
# Table of Contents

Abstract ................................................................. ii
Dedication ................................................................. ii
Acknowledgments ......................................................... iv
Vita ........................................................................ v
List of Tables ........................................................... ix
List of Figures ........................................................... x

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Introduction .................................................. 1</td>
</tr>
<tr>
<td></td>
<td>1.1 Background ................................................ 1</td>
</tr>
<tr>
<td></td>
<td>1.2 Raven’s Progressive Matrices .............................. 4</td>
</tr>
<tr>
<td></td>
<td>1.3 Statistical and Computational Models .................... 6</td>
</tr>
<tr>
<td></td>
<td>1.4 Eye tracking analysis ...................................... 7</td>
</tr>
<tr>
<td></td>
<td>1.5 Verbal protocol analysis .................................. 9</td>
</tr>
<tr>
<td></td>
<td>1.6 Raven and learning ......................................... 10</td>
</tr>
<tr>
<td></td>
<td>1.7 Experimental idea .......................................... 12</td>
</tr>
<tr>
<td>2</td>
<td>Experiment ....................................................... 13</td>
</tr>
<tr>
<td></td>
<td>2.1 Experiment .................................................. 13</td>
</tr>
<tr>
<td></td>
<td>2.1.1 Participants ............................................. 13</td>
</tr>
<tr>
<td></td>
<td>2.1.2 Stimuli ................................................... 13</td>
</tr>
<tr>
<td></td>
<td>2.1.3 Task ..................................................... 16</td>
</tr>
<tr>
<td></td>
<td>2.1.4 Procedure ............................................... 17</td>
</tr>
<tr>
<td></td>
<td>2.1.5 Apparatus ............................................... 18</td>
</tr>
<tr>
<td></td>
<td>2.1.6 Eye Movement Analysis .................................. 19</td>
</tr>
<tr>
<td></td>
<td>2.2 Results ..................................................... 21</td>
</tr>
<tr>
<td></td>
<td>2.2.1 Test Set Characteristics ................................ 21</td>
</tr>
<tr>
<td></td>
<td>2.2.2 Learning effects ........................................ 21</td>
</tr>
<tr>
<td></td>
<td>2.2.3 Components and APM score prediction .................... 23</td>
</tr>
<tr>
<td>CHAPTER</td>
<td>PAGE</td>
</tr>
<tr>
<td>---------------------------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>A Adapted Raven and Verbal Protocol Instructions (Raven, Court, &amp; Raven, 1998; Ericsson &amp; Simon, 1993)</td>
<td>37</td>
</tr>
<tr>
<td>B EyeLink 1000 Eye Tracker</td>
<td>42</td>
</tr>
</tbody>
</table>
## List of Tables

<table>
<thead>
<tr>
<th>TABLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Table showing rule agreement across the items in test sets 1 and 2 (11/14 trials). The rules types are classified according to Carpenter, Just, &amp; Shell, 1990 including pairwise progression (PP), figure addition or subtraction (A/S), distribution of three values (D3), distribution of two values (D2), and unique items (U).</td>
</tr>
<tr>
<td>2.2</td>
<td>R-squared and leave-one-out cross validated R-squared cumulative score prediction comparison for the SR/PCA and other proposed eye movement and latency measures. Regression model 1 was an attempt to replicate the results of Vigneau, Caissie, &amp; Bors, 2006 in which these three predictors accounted for $R^2=0.51$.</td>
</tr>
<tr>
<td>B.1</td>
<td>Table showing how the sample model changes based on the sample rate. The experiment in this text used the 19 sample model for analysis in DataViewer.</td>
</tr>
</tbody>
</table>
# List of Figures

<table>
<thead>
<tr>
<th>FIGURE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 Radex model showing the intercorrelations between different ability measures (Snow, Kyllonen, &amp; Marshalek, 1984).</td>
<td>3</td>
</tr>
<tr>
<td>1.2 Example of the Raven problem format and Area of Interest (AOI) layout. Problem and responses are shown with solid lines. Areas of interest are indicated by dotted lines. This example problem was generated by the author to protect the APM test integrity.</td>
<td>5</td>
</tr>
<tr>
<td>2.1 Shows the APM items and average accuracy (Raven et al., 1998; Unsworth &amp; Engle, 2005; Vigneau &amp; Bors, 2005) for each test item within test set 1 and test set 2.</td>
<td>14</td>
</tr>
<tr>
<td>2.2 Display layout and trial sequence. Subjects fixated for 1000 ms, then audio and eye movement was recorded during the decision phase. The problem was masked during the response phase.</td>
<td>16</td>
</tr>
<tr>
<td>2.3 Percent correct by item for test sets 1 and 2 (top) and item latency by item for test sets 1 and 2 (bottom).</td>
<td>22</td>
</tr>
<tr>
<td>2.4 The figure on the left shows the mean difference in SRs across sessions for participants that improved by 3 points or more (N=9) on session 2. The figure on the right shows the weight matrix (the beta-weighted sum of the the two best principal components) for all participants that predicted difference scores ($R^2=0.30$). The x-axis represents the sender AOI and the y-axis represents the receiver AOI.</td>
<td>22</td>
</tr>
<tr>
<td>2.5 The two components used to predict Raven scores are interpretable as a systematicity component (left) and a toggle component (right). The x-axis represents the sender AOI and the y-axis represents the receiver AOI.</td>
<td>23</td>
</tr>
</tbody>
</table>
CHAPTER 1
Introduction

Cognitive Psychologists have tried to explain what factors contribute to intelligence for more than a century. The appeal of understanding what makes one person perform more intelligently than another is obvious and has clear applied value. Yet despite the high level of motivation and over a century of research the precise nature of intelligence and intelligent performance remains elusive. Intelligence is difficult to study because it can be defined in a variety of different ways, by a variety of different behaviors across multiple task-specific environments. One promising approach is to identify tasks that correlate well with a multitude of other intelligence measures and study these more general tasks in greater depth.

1.1 Background

The idea of using “Correlational Psychology” to study intelligence was first formalized by Charles Spearman (Spearman, 1904). Intelligence tests that correlate well with multiple intelligence measures are said to be good measures of what Spearman called “general intelligence” or $g$. General intelligence is a mental ability factor which is theorized to contribute to all intelligent performance regardless of task domain. It has been observed that tests that are good measures of $g$ often share a common trait in that they require analytic problem-solving ability, or what Cattell called “fluid intelligence” (Cattell, 1963, 1971). More specifically, fluid intelligence ($gF$) refers to
our nonverbal abstract problem solving ability that does not rely heavily on previously learned episodic or semantic information. Cattell carefully distinguished fluid intelligence from “crystallized intelligence” or \( gC \). Crystallized intelligence refers to intellectual ability as a result of learned semantic information within specific spheres of knowledge (e.g. historical knowledge). Tests of general knowledge, vocabulary, and Verbal IQ are frequently used as measures of crystallized intelligence. There have also been a variety of intelligence tests and psychometric tasks that have been shown to measure general intelligence, some of these tasks include letter series (Thurstone, 1938), verbal analogies (Terman, 1950; Gick & Holyoak, 1983), the Wechsler Adult Intelligence Scale (Wechsler, 1955), Raven’s Progressive Matrices (Raven, 1962), necessary arithmetic operations (J. W. French, Ekstrom, & Price, 1963), letter and number analogies (D. R. Hofstadter, 1984), paper folding (J. W. French et al., 1963), and surface development tasks (J. W. French et al., 1963). Some of these tasks are also good measures of fluid ability such as Raven’s Progressive Matrices, the WAIS matrix reasoning subtest, letter & number analogies, paper folding and surface development.

This raises an important question of what method should be used to represent the correlations between these different ability measures? There have been several different methodological techniques used to attempt to map different tasks to specific or general abilities such as factor-analysis (Thurstone, 1938; P. E. Vernon, 1950; Cattell, 1963; Horn, 1976), hierarchical clustering (Sattath & Tversky, 1977) and multidimensional scaling (Guttman, 1954, 1970; Snow et al., 1984). The two most prominent models of ability organization are the Hierarchical Model and the Radex Model. The aptly named Hierarchical Model has \( g \) is at the top of the hierarchy, crystallized ability \( (gC) \), fluid ability \( (gF) \), and spatial ability \( (gV) \) as the factors of \( g \), with each of these factors broken down into even more test-specific factors (e.g. visual number span, auditory letter span) at the foundational level of the hierarchy. The exact structure
and what factors are specified varies under different schemes, for instance Vernon breaks $g$ into only two factors: verbal-educational and practical-mechanical-spatial factors (P. E. Vernon, 1950). The interpretation of the Hierarchical Model is that as tasks become more varied they become better measures of $g$. Multidimensional scaling has been shown to be mathematically equivalent to the hierarchical factor model, but arguably offers a clearer representation of what tests correlate highly with $g$ and how individual tasks relate to each other (Marshalek, Lohman, & Snow, 1983; Snow et al., 1984). The Radex model offers a clear representation in which the closer the
item is to the center of the radex, the smaller its distance to all the other tasks, the higher the task complexity, and the more closely the task is thought to be associated with \( g \). As indicated in Figure 1.1, the Radex model argues for Raven’s Progressive Matrices test as an ideal environment to explore what factors affect and contribute to \( g \), and more specifically \( gF \).

It should be noted that many dispute the theoretical claims associated with \( g \) (Thurston, 1947; Guttman, 1955; Cronbach & Gleser, 1957; Sternberg, 1985) and as was discussed earlier defining and measuring “intelligence” has proved itself to be a challenging enterprise since Spearman. However, Raven performance does correlate with a variety of other ability metrics and is a culture-fair, novel, relational environment. Perhaps, a less controversial claim is that there do seem to be certain tasks like Raven that are more closely associated with fluid mental ability than others. Specifically, it has been noted that many fluid tasks involve an analogy-making process which has been widely argued to be at the heart of intelligence (D. Hofstadter, 1995; D. R. Hofstadter, 2001; Holyoak & Thagard, 1995; Halford, 1992; R. M. French, 2002). It is with all of these views in mind that I wish to gain a better understanding of how the analogy-making process works in the Raven environment and what implications it might have for our understanding of intelligent performance at large.

1.2 Raven’s Progressive Matrices

Raven’s Progressive Matrices were originally developed by John C. Raven in 1936, as part of his Master’s thesis (Raven, 1936). More specifically, Raven’s Progressive Matrices were developed as a way to measure what Spearman called *eductive* ability, or the ability to make sense out of novel problem environments by forming high-level representations of task environments (Spearman, 1927). The basic Raven problem layout consists of a 3x3 problem matrix in which the lower right corner is missing
Figure 1.2: Example of the Raven problem format and Area of Interest (AOI) layout. Problem and responses are shown with solid lines. Areas of interest are indicated by dotted lines. This example problem was generated by the author to protect the APM test integrity.

and the task is to select the correct answer from multiple options presented at the bottom of the page (See Figure 1.2). The Advanced Progressive Matrices test (APM) was published as an ability test for adults and adolescents of above average intelligence in 1947 (Raven et al., 1998). The current version of the APM consists of two sets of items, Set I contains 12 practice problems and Set II contains 36 test items. Raven tests have been used widely over the last 70 years in areas ranging from educational achievement and clinical research to military and corporate ability assessment (Matthews, 1988; Evans & Marmorston, 1964; Flynn, 1998; Furnham, 2008).
1.3 Statistical and Computational Models

Studying Raven’s Progressive Matrices has inspired several different computational models of Raven task performance. These models can be divided into two broad categories: mental speed models (Jensen, 1987; Detterman, 1987; P. A. Vernon, 1987; Neubauer, 1990; Verguts, Boeck, & Maris, 2000) and rule-based models (Carpenter et al., 1990; Lovett, Forbus, & Usher, 2007, 2010). Mental speed models focus on information processing speed as a critical determinant of task performance. A recent meta-analysis of mental speed studies (N=53,542) found significant negative correlations between general intelligence and reaction time ($r=-0.26$), fluid intelligence and reaction time ($r=-0.21$), and crystallized intelligence and reaction time ($r=-0.17$) suggesting that mental speed is indeed an important component of intelligent performance (Sheppard & Vernon, 2008). One of the most well specified mental speed models is the Generation Speed Model (Verguts et al., 2000). In the Generation Speed Model, the rate of rule sampling is suggested as a primary factor that contributes to Raven performance. The model is tested behaviorally by showing that performance on a rule generation speed task (in which subjects list as many possible rules as they can) is correlated with APM performance ($r=0.43$).

The type and number of relations and the ability to hold relations in working memory while performing goal management are the primary performance determinants in one of the most prominent rule-based models of Raven performance (Carpenter et al., 1990). Two computational models of Raven performance were developed: FAIRRAVEN and BETTERRAVEN. The models were based on the analysis of eye fixations, verbal protocols, error patterns, and the identification of six basic rules that allow the solution of most APM problems. Both computational models were developed in the Concurrent Activation-Based Production system and consist of correspondence finding, rule induction/generalization, and working memory sub-systems. FAIRRAVEN
performance was shown to be equivalent to median subject performance on the APM, while BETTERRAVEN (with the addition of a dedicated goal management subsystem) was shown to perform like one of the best subjects in their sample. Two limitations of the models are that the Raven problems and the six identified rules are hand-coded into the production system. One recent attempt to address the limitations of the seminal Carpenter et al. model has been proposed by Lovett, Forbus, and Usher (2007, 2010). To address the first limitation, the authors propose a model that utilizes CogSketch (Forbus, Usher, Lovett, Lockwood, & Wetzel, 2008), which is a sketch understanding system in which the user inputs labeled objects which CogSketch can then use to compute spatial relations between these objects (e.g. spatial rotation). The second limitation is addressed using the Structure-Mapping Engine (SME, Falkenhainer, Forbus, & Gentner, 1989). Using SME the model computes patterns of variance between objects within the first two rows of the problem matrix and uses these correspondences to generalize an answer to the third row. The model uses two general processing strategies rather than hardwired RPM rules: a “Difference strategy” in which differences between adjacent pairs of images in a row are analyzed and a “Literal strategy” in which what is common to each image in a row is extracted and used to search for solutions. The model performance on the SPM sets B-E was 44/48 and the four missed problems were within the top six most difficult for human SPM test takers. The advancement of these models call for more research on how human observers actually process Raven matrices.

1.4 Eye tracking analysis

Given the advancements in eye tracking technology in the last 20 years, studying complex information processing strategies on Raven items is behaviorally viable. Not only
can eye tracking data increase our understanding of how humans process novel environments, but can also inform the algorithms of computational models of intelligent performance. Eye tracking data has been used to study strategy use in arithmetic (Suppes, Cohen, Laddaga, Anliker, & Floyd, 1983; Suppes, 1990; Green, Lemaire, & Dufau, 2007), mental rotation (Just & Carpenter, 1985), diagrammatic problem solving (Hegarty, 1992; Yoon & Narayanan, 2004) and visual analogies (Grant & Spivey, 2003). Eye tracking analysis has also been used to look at information processing on Raven items (Dillon, 1985; Carpenter et al., 1990; Vigneau et al., 2006).

One of the first studies to clearly demarcate two analogical processing strategies was Bethel-Fox, Lohman, and Snow (1984). In this study subjects completed geometric analogy problems in a format A:B as C:? within a 2 and 4 forced-choice design. The geometric analogy problems consisted of basic shapes such as lines, crosses, circles, squares, and triangles that were transformed by various combinations of halving, size change, doubling, rotation, and/or reflection. The authors identified two dominant processing strategies by analyzing eye tracking data: constructive matching and response elimination. An observer using a constructive matching strategy forms an idealized answer based primarily on problem information and then matches the idealized answer to the solution options. An observer using a response elimination processing strategy actively compares the problem and solutions to attempt to find an appropriate match. Constructive matching is more likely to be used by high $g_F$ individuals and on less difficult problems. Response elimination is more likely to be used by low $g_F$ individuals and on more difficult problems. It is important to point out that strategy switching dependent on ability level and difficulty was observed.

Constructive matching and response elimination strategies have also been identified on Raven APM item performance (Carpenter et al., 1990; Vigneau et al., 2006). Vigneau et al. (2006) replicated the strategy findings of Bethel-Fox et al. (1984) using
more advanced eye tracking technology and statistical methodology on Raven APM items. Participants in this study showed qualitative differences in how they inspected the problem matrix. High performance individuals inspected each cell of the matrix, while low performance individuals spent more time processing information adjacent to the missing cell. Contrary to previous findings, they also found substantially less within-subject strategy shifting.

1.5 Verbal protocol analysis

Verbal reports can also be used to study how information is processed and transformed during problem solving tasks (Ericsson & Simon, 1980; J. R. Anderson, 1987). As Ericsson and Simon argue the concurrent “think aloud” method, in which subjects verbalize their thoughts while performing a task, provides real-time insight into the subjects’s problem-relevant cognitive processes (Ericsson & Simon, 1984, 1993). The basic assumption is that the concurrent think-aloud method provides direct access to working memory (both short-term memory and information transferred from long-term memory to short-term memory) while subjects proceed through various information processing stages (Ericsson & Simon, 1993). Verbal protocol analysis has been used to study cognitive processing in a wide range of areas such as decision making (Montgomery & Svenson, 1989; M. J. Anderson & Potter, 1998), second-language learning, (Faerch & Kasper, 1987), text comprehension (Laszlo, Meutsch, & Viehoff, 1988), IQ testing (Rowe, 1985), memory (Bellezza, 1986; Ericsson, 1988), human factors (Deffner, 1990), software engineering (see Hughes & Parkes, 2003 for review), strategic resource-allocation (Ball, Langholtz, Auble, & Sopchak, 1998), education (Chi, Bassok, Lewis, & Reimann, 1989; Trudel & Payne, 1995), and artificial intelligence (Conati & VanLehn, 2000).

Two of the main methodological challenges involved in using the think-aloud
method is determining whether it will result in reactivity (impact natural processing) and be complete (provide access to all the relevant cognitive processing) (Wilson, 1994). Raven is a particularly interesting domain for verbal protocol analysis as it requires the processing of non-verbal information due to its consistent non-verbal content of regular and irregular shapes and patterns. Studies that compare think-aloud to silent conditions on Raven (Rhenius & Heydemann, 1984; Heydemann, 1986) and similar figural analogy problems (Deffner, 1983, 1989) have two consistent findings: subjects in the think-aloud and silent solution conditions show no reliable difference in accuracy, and subjects in the think aloud condition are slower on easier items than their silent solution counterparts. Therefore, there is evidence that with the exception of reaction time measures the think-aloud method is not reactive on visual analogy problems despite their decidedly non-verbal format. The issue of completeness is not nearly as clear. The processing of geometric analogies is not well understood, therefore, it is unclear how much processing is automatic or unconscious. However, there is evidence that the think-aloud method is well suited to complex problem-solving that involves multiple conscious processing steps. Moreover, it is encouraging that verbal protocols have previously been used to assist in understanding and modeling Raven task performance (Carpenter et al., 1990). So while it is not as clear cut whether Ericsson and Simon’s think-aloud procedure is complete on Raven, there is some positive evidence that it may be complete given that verbal reports have been useful in conjunction with eye tracking in the past.

1.6 Raven and learning

There are relatively few studies looking at how Raven processing changes with learning over repeated testing sessions (Denney & Heidrich, 1990; Bors & Vigneau, 2003). Denney & Heidrich, 1990 examined whether there were practice effects for young,
middle-aged, and elderly adults on Raven items from the SPM and APM tests. Training in this study consisted of strategy modeling in which the participant listened to the experimenter verbalizing her strategy as she worked through three Raven problems, and then the participant used the same strategy to solve three similar problems. The increases due the strategy modeling technique were more modest (approximately 1 point) than those observed on longer multi-session training done with other fluid tasks (approximately 2 points), but were obtained using substantially fewer training trials. However, only having measures of accuracy and reaction time made it difficult to say anything concrete about how information processing strategy changed at post-test.

In a more recent examination, individual differences and correlations between similar Raven items were used to try to understand what is learned through practice on the APM test (Bors & Vigneau, 2003). Participants completed the APM (12-item practice set and 36 item test) on three separate occasions with intervals of approximately 45 days between sessions. The results showed improvement across sessions, but did not offer any insight into what this improvement represented in terms in strategic processing. It did tell us is that individual subject improvement was highly variable with some subjects improving greatly and even some getting worse across sessions, but the rank ordering of individuals scores was stable across sessions. There was a high degree of item switching (missing items that had previously been answered correctly and vice versa) between testing sessions and the number of items left unanswered did not account for the improvements, which suggests that the subjects did not improve due to remembering specific problems. This study is evidence that something is indeed changing with increased exposure. However, it remains unclear what contributed to the improvement in performance and it is an
open question as to what these improvements actually represent in terms of strategic information processing or analogical processing mechanisms.

1.7 Experimental idea

Therefore, while there have been studies looking at processing strategy and studies looking at learning, there have not been any studies that examine how learning affects strategic processing on the APM. More importantly, the interaction between learning and strategy could shed light on whether shifts in strategic processing account for the observed improvement in earlier practice effect studies, which could in turn could inform computational performance models. The present study explored how learning affects strategy use on the APM. A novel methodology was used that represents the spatio-temporal nature of eye movements using successor representations (Dayan, 1993) and assesses individual differences in strategy across days via principal component analysis.
2.1 Experiment

2.1.1 Participants

Thirty-five university students with normal or corrected-to-normal vision participated in the study. Participants were paid 6 dollars per hour plus a bonus contingent upon their accuracy.

2.1.2 Stimuli

Two computerized test sets were created using the items taken from the APM. Both test sets were short-form 14 item versions of the APM that were chosen to represent a wide variety of relations (Carpenter et al., 1990) and difficulty levels (Raven et al., 1998). Test set 1 was composed of the following 14 APM Set II items: 2, 4, 6, 9 10, 11, 16, 17, 19, 21, 23, 24, 26, and 29. Test set 2 was composed of another unique set of 14 APM Set II items: 1, 3, 5, 7 , 12, 13, 14, 15 18, 20, 22, 25, 27, and 28. Short-form versions of the APM have been found to dramatically reduce administration time while maintaining many of the desirable psychometric properties of the full APM (Arthur & Day, 1994; Bors & Stokes, 1998; Caffarra, Vezzadini, Zonato, Copelli, & Venneri, 2003). The test items were ordered as shown in Figure 2.1 based on an average of the accuracy data reported in the APM manual (Raven et al., 1998) and other recent findings (Unsworth & Engle, 2005; Vigneau & Bors, 2005). For the
Figure 2.1: Shows the APM items and average accuracy (Raven et al., 1998; Unsworth & Engle, 2005; Vigneau & Bors, 2005) for each test item within test set 1 and test set 2.

purposes of eye tracking each APM item was converted into a digital vector image and digitally altered in Adobe Photoshop (Adobe, 2005). The scale of the matrix and responses was increased 78% to increase overall eye tracking resolution without altering the aspect ratio of the original APM items. In order to allow individual interest areas for each response to be accurately tracked, the gaps between response items were increased from 0.22° to 2.30°, but the responses still appeared as two rows of four choices across the bottom of the screen (see Figure 1.2). The gap between the matrix and response options was also increased from 1.81° to 3.05°. No other alterations were made to the scanned images. The changes successfully allowed individual cells within the problem matrix and response area to be tracked accurately. Existing studies of computerized versions of Raven tests do not report any significant
differences in performance on computerized vs. paper and pencil versions of the test (Williams & McCord, 2006; Arce-Ferrer & Guzman, 2009).

<table>
<thead>
<tr>
<th>Trial</th>
<th>Test 1 item</th>
<th>Rule type</th>
<th>Test 2 item</th>
<th>Rule type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>PP</td>
<td>1</td>
<td>D3</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>PP</td>
<td>3</td>
<td>PP</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>PP</td>
<td>5</td>
<td>PP</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>A/S</td>
<td>7</td>
<td>A/S</td>
</tr>
<tr>
<td>5</td>
<td>11</td>
<td>A/S</td>
<td>12</td>
<td>A/S</td>
</tr>
<tr>
<td>6</td>
<td>10</td>
<td>PP</td>
<td>14</td>
<td>PP</td>
</tr>
<tr>
<td>7</td>
<td>16</td>
<td>A/S</td>
<td>15</td>
<td>A/S</td>
</tr>
<tr>
<td>8</td>
<td>17</td>
<td>D3</td>
<td>13</td>
<td>D3</td>
</tr>
<tr>
<td>9</td>
<td>19</td>
<td>U</td>
<td>18</td>
<td>U</td>
</tr>
<tr>
<td>10</td>
<td>21</td>
<td>D3</td>
<td>20</td>
<td>A/S</td>
</tr>
<tr>
<td>11</td>
<td>23</td>
<td>D2</td>
<td>22</td>
<td>D2</td>
</tr>
<tr>
<td>12</td>
<td>24</td>
<td>PP</td>
<td>25</td>
<td>PP</td>
</tr>
<tr>
<td>13</td>
<td>26</td>
<td>PP</td>
<td>27</td>
<td>D3</td>
</tr>
<tr>
<td>14</td>
<td>29</td>
<td>D3</td>
<td>28</td>
<td>D3</td>
</tr>
</tbody>
</table>

Table 2.1: Table showing rule agreement across the items in test sets 1 and 2 (11/14 trials). The rules types are classified according to Carpenter et al., 1990 including pairwise progression (PP), figure addition or subtraction (A/S), distribution of three values (D3), distribution of two values (D2), and unique items (U).
2.1.3 Task

The timeline of each test trial is illustrated in Figure 2.2. Each test trial began with a brief alert sound that let the observer know they should fixate the cross on the screen. The fixation screen proceeded every trial to ensure eye-tracking accuracy of 0.5° or better on every trial. If the observer successfully fixated the cross for 1000-msec the problem stimulus would appear, otherwise a re-calibration was performed. The problem stimulus consisted of the problem matrix, 8 responses as two rows of four choices, and a mouse cursor in a box between the problem matrix and response options. The problem remained on screen until the observer moved the cursor out of the box at which point the problem was masked and only the responses remained on screen. Audio started recording when the problem appeared and ended when observer moved the cursor out of the box. No trial or test time-limit was enforced.
as is recommended by the APM manual due to studies showing the high correlation between timed and untimed testing conditions (P. A. Vernon, Nador, & Kantor, 1985; P. A. Vernon & Kantor, 1986; Sternberg, 1986).

2.1.4 Procedure

Each session began with a verbal instruction (See Appendix A) that consisted of the experimental procedure for how to fixate prior to each trial, how to select an answer using the mouse, instructions on the APM items that followed the APM manual guidelines (Raven et al., 1998), and verbal protocol instructions (Ericsson & Simon, 1993). The APM manual requires that you walk through a demo problem with the subject to ensure they understand the task. A simple pattern completion problem was taken from the significantly easier Standard version of Raven’s Progressive matrices (SPM item A01) to avoid unintended exposure to more complex relations appearing in the APM test items. The verbal protocol instructions followed the basic instruction script in the appendix of Protocol Analysis: Verbal Reports as Data (Ericsson & Simon, 1993). Participants completed unrelated examples such as adding while they thought aloud until the experimenter was satisfied that the participant was comfortable with the procedure. During the verbal protocol instructions participants were also told that if they are silent for 10 seconds they would receive a message to “Please keep talking” and this was to encourage them to continue to think aloud.

To ensure participants understood the the task they also performed two demo trials on the computer while being tracked and having their voice recorded. The first demo trial that appeared was the same one used in the paper and pencil instructions and ensured the subjects understood the mechanics of the experimental procedure (i.e. fixating when they hear the alert beep and moving the cursor to select their answer). The second demo problem was one they had not seen (SPM item A03) and
allowed the experimenter to assess verbal protocol on a related problem. Then, if
the subject had no questions, the pre-test began and the experimenter left the room.
The experimenter only returned if the eye tracker needed to be re-calibrated due to
a failure during the fixation phase.

2.1.5 Apparatus

The APM items were presented on a 21” NEC Accusync 120 color CRT driven by a
2.66 GHz Intel iMac computer. The monitor was the only light source in the room
and was viewed binocularly with the natural pupil from a chin rest located ≈92 cm
away. At that distance, one degree of visual angle spanned ≈43 pixels (1024 × 768
resolution).

Monocular Eye movements were recorded using an Eyelink 1000 desktop eye
tacker (SR Research, 2006) at a sampling rate of 1000 Hz. Nine-point calibration
was performed using the pupil and marker procedure within the Eyelink software
and resulted in an average accuracy of 0.5° or better. Saccades were detected using
a EL1000 19-sample model with a velocity threshold of 30°/sec. and an acceleration
threshold of 8000°/sec. The experiment was created and controlled using the Exper-
iment Builder software (SR Research, 2010b). Data was analyzed using Data Viewer
(SR Research, 2010a) and Matlab (The MathWorks, 1999).

Verbal reports were recorded using an E-MU 0202 USB 2.0 audio interface and
Shure beta 58A supercardioid dynamic microphone. The recorded protocols were
segmented using Logic Studio audio editing software (Apple, 2009) and the audio
segments were then coded using the Computer-Aided Protocol Analysis System (CA-
PAS 2.0) developed by Robert Crutcher (Crutcher, 2007).
2.1.6 Eye Movement Analysis

The area of interest analysis allowed for a wide array of fixation and saccade metrics. Part of the analysis focused on traditional AOI measures that are thought to indicate strategic processing such as proportional dwell time, toggle rate, and matrix time distribution index. These traditional AOI measures are compared to a novel method of analysis that analyzes sequential fixation patterns within the problem via successor representations (Dayan, 1993) and standardized principal component analysis.

Ten consistent areas of interest (AOIs) were defined on each item. The first nine AOIs were the nine cells of the problem matrix (top row = 1 2 3, middle row = 4 5 6, bottom row = 7 8 9) and the final area “R” consisted of the entire response area (See Figure 1.2). A sequence of eye movements was defined as the sequential pattern of fixations across the 10 different areas by a participant on a given trial. Fixation events that did not fall in one of these 10 areas were removed from the sequences (<1% of total fixations). Each sequence had three distinct phases orientation, solution, and answer selection. The answer selection phase consists of approximately the last 5% of trials, however, as trials become more difficult the orienting phase increases in length. In order to focus on the solution phase, sequences that exceeded the median sequence length by 50 fixations or more had 20% of total sequence length clipped from the beginning of the sequence. To control for answer selection all trials had 5% clipped from the end of the sequence.

Statistical regularities in these complex patterns of fixations were captured using successor representations (SRs, Dayan, 1993). Unlike a traditional transition-probability analysis where each cell of the 10 x 10 matrix represents the frequency of making a single saccade from one AOI to another, the SR uses temporal difference learning to incrementally strengthen the weights of multiple cells based on both recent and future transitions. The result is a matrix representation that integrates over
multiple time steps to estimate the expected discounted number of future fixations a location \( j \) given a current fixation at location \( i \). Calculating the SR of a sequence requires two simple calculations: a prediction error calculation and a temporal difference weight update rule. At the beginning of each trial the SR (call it \( M \)) is a 10 x 10 matrix of zeros and \( I \) is the identity matrix of \( M \). The prediction error, given a state \( i \), is equal to the the identity vector of the next state plus the product of gamma and the SR vector of the next state minus the SR vector of the current state (See Equation 2.1). The weights are updated via an incremental temporal difference learning rule calculated as the current SR representation plus the learning rate (alpha) multiplied by the outer product of the prediction error and the transpose of the identity vector of the current state (See Equation 2.2).

\[
PE = (I_{i+1} + \gamma M_{i+1}) - M_i
\]  
(2.1)

\[
\Delta M = \alpha (PE_i \times I^T_i)
\]  
(2.2)

SRs were generated for each trial for every participant and averaged across trials to generate a mean SR for each participant. A standardized principal component analysis (SPCA) was performed by applying a singular value decomposition to the z-transform of the mean SR data and the z-transformed data was projected onto the orthogonal coordinate axes defined by the principal components. The principal components were then used to predict participant scores. A leave-one-out cross validation that optimized the SR parameters by maximizing \( R^2 \) prediction using a Nelder-Mead simplex algorithm. The cross validation results indicated that the same two components were chosen on all 35 cross validation runs and showed consistent parameter ranges across the 35 runs (gamma=0.22±0.03 and alpha=0.23±0.02). For simplicity’s sake, the results reported below optimized the SR parameters based on \( R^2 \) for all 35 participants.
2.2 Results

2.2.1 Test Set Characteristics

Scores on test set 1 ranged from 4 to 14 with a mean accuracy of 9.58 (S.D.=2.29). Scores on test set 2 ranged from 6 to 13 with a mean accuracy of 10.52 (S.D.=2.01). The mean item latency was 95 seconds (S.D.=58) for test set 1 and 85 seconds (S.D.=56) for test set 2. The degree of variability in item latency reflects moderate individual differences in item completion time, particularly among late test items. There was a linear increase in item latency as each test progressed (r=0.90 test 1, r=0.85 test 2), which was expected as item difficulty increases as each test progresses. The difference in mean performance between test set 1 and test set 2 at pretest was not significant (t\(_{33}\)= -0.80, p=0.43). Due to the test similarity across relation type, item accuracy, item latency, and difficulty, test version was collapsed over in the subsequent data analysis (See Figure 2.3).

2.2.2 Learning effects

In agreement with earlier studies (Denney & Heidrich, 1990; Bors & Vigneau, 2003) we also observed a significant learning effect across sessions (mean gain= 1.52, t(34)=3.481, p<.001). In an attempt to explain the increase in score across sessions we altered our methodology to use the difference in SRs (session 2 SRs - session 1 SRs) to predict the difference in scores (session 2 score - session 1 score). The results showed that components which captured increased scanning systematicity on the second session could explain a significant portion of variance in the difference scores (\(R^2=0.30\)). The participants which showed the largest learning effect also had the greatest increase in SR systematicity, which informed the principal components of the PCA (See Figure 2.4).
Figure 2.3: Percent correct by item for test sets 1 and 2 (top) and item latency by item for test sets 1 and 2 (bottom).

Figure 2.4: The figure on the left shows the mean difference in SRs across sessions for participants that improved by 3 points or more (N=9) on session 2. The figure on the right shows the weight matrix (the beta-weighted sum of the the two best principal components) for all participants that predicted difference scores ($R^2=0.30$). The x-axis represents the sender AOI and the y-axis represents the receiver AOI.
2.2.3 Components and APM score prediction

The top two components from the PCA each explained large proportions of variance in cumulative scores (sum of session 1 and session 2 scores) and had clear interpretations. As is evident in Figure 2.5 the first component had an obvious diagonal "box" structure which accounted for the largest proportion of the variance in Raven scores ($R^2=0.31$, $\beta = 0.48$). The 3 x 3 diagonal boxes are indicative of the benefits of systematically scanning within a given row of the problem matrix as opposed to haphazard scanning or column-wise scanning. The red areas below each box indicated systematic integration as participants moved from row to row. The second component’s dominant feature is the solid blue line across the response area ($R^2=0.25$, $\beta = 0.95$). The solid blue area is interpreted as an "anti-toggle" component. That is, participants that make fewer toggles from each cell of the problem matrix to the response area have higher scores than participants that toggle more frequently. Together these two components explained 56% of the variance in cumulative Raven scores (See Table 2.2).
Table 2.2: R-squared and leave-one-out cross validated R-squared cumulative score prediction comparison for the SR/PCA and other proposed eye movement and latency measures. Regression model 1 was an attempt to replicate the results of Vigneau et al., 2006 in which these three predictors accounted for $R^2=0.51$. 

<table>
<thead>
<tr>
<th>Measure</th>
<th>$R^2$</th>
<th>cv$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SR/PCA</td>
<td>0.56</td>
<td>0.51</td>
</tr>
<tr>
<td>Regression Model 1 (MTDI, TR, IL)</td>
<td>0.17</td>
<td>0.04</td>
</tr>
<tr>
<td>Proportional matrix dwell time (PMDT)</td>
<td>0.17</td>
<td>0.09</td>
</tr>
<tr>
<td>Toggle rate on easy items (TR)</td>
<td>0.12</td>
<td>0.04</td>
</tr>
<tr>
<td>Item latency on easy items (IL)</td>
<td>0.11</td>
<td>0.04</td>
</tr>
<tr>
<td>Matrix time distribution index (MTDI)</td>
<td>0.02</td>
<td>0.01</td>
</tr>
</tbody>
</table>
CHAPTER 3
Discussion

The novel SR/PCA methodology produced new results both in terms of successful score prediction as well as new insight into the influence of strategy on Raven performance. The successor representation and principal component analysis extracted the important underlying structure from within complex patterns of sequential eye movements and used individual differences in this structure to predict APM scores with an unprecedented level of success. Moreover, the principal components produced by the PCA were readily interpretable and offered new insights into strategic processing and the learning effect on Raven. These are important developments in understanding Raven’s Progressive Matrices and what accounts for individual differences in performance.

The anti-toggle component replicated earlier findings that suggested toggling is indicative of a non-optimal strategy on Raven. The SR representation and SPCA was able to predict a larger proportion of variance in Raven scores than traditional measures (such as number of toggles or toggle rate) suggesting it may be a substantially more sensitive measure of toggling. However, it remains unclear whether increased toggling is a result of an initial response elimination strategy at the beginning of a problem, as a fall back strategy when participants are struggling to find relations, or both. Further research is needed to make this important distinction.

The systematicity component was a novel finding and arguably provides the most detailed picture of Raven performance and strategic processing. The systematicity
component demonstrates the importance of strategically processing rows in a highly systematic fashion. Within rows there was also evidence that integrating cell information is more successful if it is attained by scanning adjacent cells (1→2, 2→3, 3→2) as opposed to across cells (1→3, 4→6). This would imply that row scanning (particularly adjacent cell scanning within rows) is more likely to generate relational insight, which would conform to previous findings that perceptual-motor patterns can increase the likelihood of rule insight (Grant & Spivey, 2003). This is an exciting new finding on Raven which lends new support to the theory that successful Raven solvers use a constructive matching strategy, while providing a more detailed illustration of what constructive matching entails.

In addition, by adjusting our methodology to analyze the differences in SRs for session 1 and session 2 we revealed an increase in SR systematicity on session 2 relative to session 1, particularly among high learners. In terms of strategy this suggests that some participants are not only altering their strategies across trial difficulty, but that they are also altering their strategy across sessions. One interesting interpretation of this finding is that with increased Raven exposure a significant portion of participants are “learning” to use a more systematic strategy even in the absence of feedback.

Finally, the success in terms of prediction and interpretation on Raven argues for the usefulness of the novel SR/PCA methodology in a much broader context. It is likely that this method could prove useful in identifying information processing strategies of experts and novices in any complex task environment that has distinct areas of interest. One can easily imagine this same method being applied to other abstract, rule governed environments such as chess or the Tower of Hanoi, as well as more applied areas like identifying successful and unsuccessful information processing strategies for landing a plane or driving a car.
References


Guttman, L. (1954). A new approach to factor analysis: The radex. In P. F. Lazerfield...


Vernon, P. A., & Kantor, L. (1986). Reaction time correlations with intelligence test scores obtained under either timed or untimed conditions. *Intelligence, 10*, 315–330.


Appendix A

Adapted Raven and Verbal Protocol Instructions (Raven et al., 1998; Ericsson & Simon, 1993)

Task and Eye Tracking Instruction Section

**SAY:** In this experiment we are interested in looking at what peoples eyes do while they solve problems. So during the experiment you will be presented with a series of puzzles. There will be two different tasks that you have to complete while participating in this experiment.

**SAY:** First, you will need to look at a cross at the center of the screen before each problem is presented on the screen. You will know when you need to do this because you will see a screen like this (show fixation piece of paper) and you will hear a beep sound. Keep staring at this cross until the problem appears. This step is required to make sure the eye tracker is tracking you properly throughout the experiment. If it is not tracking you properly an instruction will appear on screen that tells you to get the experimenter, so if that happens, please come and get me and I will come back and perform a recalibration.

**DO:** Point to the cross to show subjects where they are supposed to look.

**SAY:** After you look at the cross, then the actual problem we want you to solve will be presented. All of the problems have this basic form...

**DO:** Show subject the example Raven problem 1.
**SAY:** The top part of the problem is a pattern with a piece cut out of it.

**DO:** Point to the blank space in the upper figure.

**SAY:** Each of the figures shown here

**DO:** Move finger across eight options.

**SAY:** is the right shape to fill the space, but they are not all the correct pattern. Only one of them is the pattern that came out of the space. Point to the one which, if it was put back, would complete the pattern both along and down.

**DO:** Move finger along the horizontal rows of dots in the upper figure and down the vertical lines, then pause in the blank space.

**SAY:** Which is the one that came out of here?

* If the person taking the test does not point to any answer, offer further explanation until he or she has grasped what is required.

* If the person taking the test chooses a wrong answer, show him or her why it is wrong and ask for another choice.

**Once they choose the correct answer**

**SAY:** Yes, that’s right. It’s the only one that completes the pattern correctly, both along and down. Once you have decided on an answer in the experiment, you will select your answer by moving the cursor out of the box and then clicking on the answer (right or left click will work). *IMPORTANT* only move the cursor out of the box when you are ready to select your answer as it causes the problem to be covered up.

**If the person makes a mistake on the example problem**

**SAY:** Look carefully at the pattern and remember that one and only one of the eight
pieces shown below is exactly right. Be sure that you choose the one that makes the pattern right, both along and down.

**If the person taking the test expresses anxiety over minute details**

**SAY:** There are no catches or tricks in this test. All the figures are correctly drawn and the options are either right or wrong.

---

**After paper and pencil demo problem**

**SAY:** The problems in the experiment are similar to those you just completed. The only difference is there are more of them and you will find that the problems get progressively more difficult.

**SAY:** Remember it is accurate work that counts ($1.00 per correct answer). Do your best to find the correct piece to complete the problem before going on to the next. But remember, in every case the next problem is harder and it will take you longer to check your answer carefully.

*After task instructions and the paper and pencil demo problem is completed successfully, move on to the Verbal Protocol Instruction Section*

---

**Verbal Protocol Instruction Section**

**SAY:** In this experiment we are also interested in what you think about as you find answers to the problems we ask you to solve. In order to do this I am going to ask you to THINK ALOUD as you work on the problems. What I mean by think aloud is that I want you to tell me EVERYTHING you are thinking from the time you first see the question until you give an answer. I would like you to talk aloud constantly from the time I present each problem until you have given your final answer to the question. I dont want you to try to plan out what you say or try to explain to me what you are saying. Just act as if you are alone in the room speaking to yourself.
It is very important that you keep talking. If you stop talking for a period of time you will receive an audio message that reminds you to “Please keep talking”. Do you understand what I want you to do?

**SAY:** Good, now we will begin with some un-related practice problems. First, I want you to:

Suggested problems

1) talk aloud while you add 76 plus 54
2) talk aloud while you multiply 24 times 4
3) talk aloud while you solve the following anagram ¡NPEPHA = HAPPEN¿
4) talk aloud while you think of as many words as possible that rhyme with beef
5) talk aloud while you try to tell me how many windows are in your parents house

**Practice problems should be given until the subject is comfortable with thinking aloud**

**SAY:** Very good. Do you have any questions? Now before we move into the eye tracking room and begin the experiment we will do one practice problem similar to the type you will be completing including the eye tracking procedure discussed earlier.

Use paper and pencil task demo sheet again except this time include verbal protocol

1) For review present Fixation sheet (Ask subject what they are supposed to do when presented with this screen?)
2) Present Raven Problem and have them think aloud while they solve it
3) Then ask them how they select their answer (to check once again to make sure they understand the procedure)

**SAY:** Great. Now that you understand the task procedure and the think aloud procedure will move into the eye tracking room and begin the actual experiment.
Once in the eye tracking room

SAY: First, we will perform a new calibration.

After calibration

SAY: It is important that you keep your head and body as still as possible so that the eye tracker can continue to track where you are looking properly.

SAY: Now I will stay here while you complete the first two demo problems to ensure you understand the task and in case you have any questions.

After subject completes first 2 demo trials

SAY: You did very well. Do you have any questions before you begin?

SAY: Okay, I will be in the next room. Remember to keep thinking aloud while you complete the problems. There will be a break halfway through the experiment. At this point feel free to take a short break; get up, stretch, rest your eyes, or go to the bathroom at this point.

Experimenter leaves the room
Appendix B

EyeLink 1000 Eye Tracker

**Saccade parser algorithm:** The EyeLink 1000 eye tracker uses a saccade detector mechanism to separate saccades from fixations. For details on the saccade detection and filtering algorithms see Stampe, 1993.

**Velocity and Acceleration:** Velocity and acceleration are computed by a weighted sum of samples, and scaled by instantaneous PPD (pixels per visual degree). The velocity filter is equivalent to averaging the instantaneous velocity of two samples (e.g. for 500 Hz):

$$V_{f[n]} = \frac{500(x_{[n+1]} - x_{[n-1]})}{2PPD_{[n]}}$$  \hspace{1cm} (B.1)

Acceleration is computed with similar filters, equivalent to the weighted sum of the instantaneous acceleration of three samples:

$$A_{[n]} = \frac{500^2(x_{[n-2]} - 2x_{[n]} + x_{[n+2]})}{4PPD_{[n]}}$$  \hspace{1cm} (B.2)

**Event definitions:**

**Saccade:** If two or more consecutive samples exceed the velocity or acceleration thresholds then a saccade is generated and continues until the start of a period of saccade detector inactivity of 20 msec.

**Blink:** A blink is defined as a period of activity in which the pupil data is missing for three or more consecutive samples.
**Fixation**: A fixation is defined as any period of samples that are not classified as a blink or saccade event.

**EyeLink 1000 sample model:**
As the sample rate increases the number of samples used for calculating velocity and acceleration must also increase due to increased noise. The EL1000 sample model automatically adjusts the number of samples used based on the sample rate.

<table>
<thead>
<tr>
<th>Sample Rate</th>
<th>Sample model size</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000 Hz</td>
<td>39 sample model</td>
</tr>
<tr>
<td>1000 Hz</td>
<td>19 sample model</td>
</tr>
<tr>
<td>500 Hz</td>
<td>9 sample model</td>
</tr>
<tr>
<td>250 Hz</td>
<td>5 sample model</td>
</tr>
</tbody>
</table>

Table B.1: Table showing how the sample model changes based on the sample rate. The experiment in this text used the 19 sample model for analysis in DataViewer.