Mechanisms of Visual Relational Reasoning

DISSERTATION

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of the Ohio State University

By
Taylor Ray Hayes, B.A., M.A.
Graduate Program in Psychology

The Ohio State University
2014

Dissertation Committee:
Alexander A. Petrov, Advisor
Roger Ratcliff
Per B. Sederberg
James T. Todd
Abstract

Humans possess an extraordinary ability to extract relational information even in completely novel task environments. What are the underlying mechanisms that make this relational extraction process possible? The presented work investigates new aspects of this question by studying how people solve visual analogy problems from a benchmark test of fluid intelligence, Raven’s Advanced Progressive Matrices (APM). Fluid intelligence is of great functional importance because it represents an important dimension of individual differences in cognition and is one of the best predictors of life outcomes such as educational and job performance. Relational reasoning was studied in a series of experiments using a rich data sources (APM test score, solution time, eye movement patterns, pupil diameter, and verbal protocols) within novel experimental paradigms. In one study we develop a novel method for extracting statistical regularities from complex sequences of eye movements (Successor Representation Scanpath Analysis, SRSA) to quantify the role of individual differences in attentional control during Raven performance. The results revealed that 41% of variance in APM score could be explained by individual differences in problem solving strategies. In a followup study SRSA was used to study APM practice effects commonly observed in the test-retest designs used in the cognitive enhancement literature. The findings revealed that APM practice effects can be explained by refinements in strategy and raised the possibility that strategy refinement may represent a potentially serious confound in cognitive training research that uses visual
training regimens (e.g., visual N-back tasks). In two followup studies relational reasoning was studied by actively manipulating the order and amount of APM problem information presented to the observer. The findings established a causal relationship between information order and relational insight and suggested participants deploy attention to one cell of the problem matrix at a time. Finally, a novel combination of pupillometry and verbal protocol analysis was used to understand how the mediation of the exploration-exploitation tradeoff contributes to individual differences in fluid intelligence. Converging evidence from primate electrophysiology and computational neural modeling have indicated that changes in exploratory versus exploitive control state may be mediated by the broad noradrenergic projections emanating from the locus coeruleus (LC). At the same time, pupil diameter has has recently emerged as a promising noninvasive proxy measure for LC activity. While Raven is well-suited for studying exploration-exploitation tradeoffs, it also posed a major obstacle to measuring pupillary response: pupil foreshortening error (PFE). PFE is caused by changes in gaze position which produce foreshortening of the pupillary image because the eye tracking camera is fixed but the eye rotates. In a supplementary study, PFE was systematically mapped using an artificial eye model and then corrected by a geometric model that expressed the foreshortening of the pupil area as a function of the cosine of the angle between the eye-to-camera axis and the eye-to-stimulus axis. The model reduced the root mean squared error of pupil measurements by 97.5% and made it possible to measure cognitive pupillary effects during Raven solution. A final study pioneered the use of pupillary response in conjunction with verbal protocol analysis to understand how the mediation of the exploration-exploitation tradeoff contributes to visual relational reasoning. The results showed an increase in pupil diameter during exploratory periods and decrease during exploitative periods consistent with prominent theories of LC function. In addition, 25% of the variance
in APM score was explained by individual differences in pupillary response during exploration. This is the first study showing evidence that noradreneric activity may mediate the exploration-exploitation tradeoff during analogical reasoning and the only study to show an exploration-related boost in pupillary response that is modulated by individual differences in fluid intelligence.
To Erika.
Acknowledgments

I owe my deepest gratitude to my advisor, Alex Petrov, that guided and encouraged me throughout my graduate career from green graduate student to doctoral candidate. I have benefited in immeasurable ways from his mentorship and I am proud to have worked in his lab. I would also like to thank the members of my dissertation committee, Roger Ratcliff, Per Sederberg, and Jim Todd, each of whom have taught me a great deal during my time as a graduate student. I am also indebted to my family who have always supported me in my academic pursuits over the years. Finally, I owe a debt of gratitude to my fiancée, Erika, for her continued support and encouragement while I completed my dissertation.
Vita

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

2011 ....................................... M.A. Cognitive Psychology, The Ohio State University

2008-Present ............................... Graduate Teaching Associate, Department of Psychology, The Ohio State University

Publications


Fields of Study

Major Field: Psychology

Specialization: Cognitive Psychology
Table of Contents

Abstract .............................................. ii
Dedication ........................................ iv
Acknowledgments ................................. vi
Vita .................................................... vii
List of Tables .................................... xii
List of Figures .................................... xiii

CHAPTER PAGE

1 Introduction .................................... 1

2 A novel method for analyzing sequential eye movements reveals strategic influence on Ravens Advanced Progressive Matrices .......................... 11
  2.1 Introduction .................................. 11
  2.2 Experiment ................................... 16
    2.2.1 Method .................................. 16
    2.2.2 Results and Discussion ................. 20
  2.3 General Discussion .......................... 28

3 Do we really become smarter when our fluid-intelligence test scores improve? ................................................ 32
  3.1 Introduction .................................. 32
  3.2 Method ....................................... 40
    3.2.1 Relational Item Scoring ................. 41
    3.2.2 Successor Representation Scanpath Analysis .......................... 44
    3.2.3 Results .................................. 47
  3.3 General Discussion .......................... 52
4 Quantifying the role of strategy and attentional scope during relational reasoning ........................................... 63

4.1 Introduction ........................................................................................................................................ 63
4.2 Experiment 1 ........................................................................................................................................ 70
  4.2.1 Method .......................................................................................................................................... 71
  4.2.2 Results and Discussion ................................................................................................................. 72
4.3 Experiment 2 ........................................................................................................................................ 75
  4.3.1 Method .......................................................................................................................................... 75
  4.3.2 Results and Discussion ................................................................................................................. 76
4.4 General Discussion ................................................................................................................................. 78

5 Mapping and correcting the influence of gaze position on pupil size measurement ................................... 81

5.1 Introduction ........................................................................................................................................ 81
5.2 Experiment 1 ........................................................................................................................................ 87
  5.2.1 Method .......................................................................................................................................... 88
  5.2.2 Results and Discussion ................................................................................................................. 89
5.3 Experiment 2 ........................................................................................................................................ 93
  5.3.1 Method .......................................................................................................................................... 94
  5.3.2 Results and Discussion ................................................................................................................. 99
5.4 General Discussion ................................................................................................................................. 104

6 Relational Foraging: Tracking exploration-exploitation trade-offs during relational reasoning .................. 115

6.1 Introduction ........................................................................................................................................ 115
6.2 Method ................................................................................................................................................ 118
  6.2.1 Pupil Data Preprocessing and Gaze Position Correction ......................................................... 120
  6.2.2 Verbal Protocol Coding Procedure ......................................................................................... 121
  6.2.3 Pupil and Verbal Protocol Analysis ......................................................................................... 122
6.3 Results ................................................................................................................................................ 124
6.4 Discussion ........................................................................................................................................... 126

7 General discussion ................................................................................................................................... 129

7.1 Future Directions ................................................................................................................................ 134

References .................................................................................................................................................. 137

Appendix A: Supplementary information for chapter 2 .................................................................................. 158

A.1 SRSA technical details .......................................................................................................................... 159
### List of Tables

<table>
<thead>
<tr>
<th>TABLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Goodness-of-fit $R^2$ and leave-one-out cross-validated $R^2_{cv}$ for predicting Raven scores from eye-movement data. The top line reports the performance of the novel method based on successor representations and principal component analysis (PCA). It is compared to some prominent dwell-time variables from the literature (Vigneau, Caissie, &amp; Bors, 2006) and to first- and second-order transition probability matrices.</td>
</tr>
<tr>
<td>5.1</td>
<td>Descriptive statistics of the deviation in pupil diameter from the geometric mean for the 9 empirical maps (3 experimental layouts x 3 pupil sizes). RMSE = root mean squared error.</td>
</tr>
<tr>
<td>5.2</td>
<td>Descriptive statistics of the deviation in pupil diameter from the geometric mean for the aggregate layout maps, the corrective multipliers generated by the parameter-free (PF) model and the model with optimized (opt.) parameters, and the corrected data (parameter-free and optimized). RMSE = root mean squared error.</td>
</tr>
<tr>
<td>5.3</td>
<td>Comparison of geometric models (parameter-free physical measurements vs. optimized parameters) for Near, Medium, and Far Layouts. All values are distances in mm. Cx, Cy, Cz are the coordinates of the camera; Sx, Sy, Sz are the coordinates of the upper left-hand corner of the screen (cf. Figure 5.3). An asterisk indicates the parameter was fixed to the physical layout measurement.</td>
</tr>
<tr>
<td>B.1</td>
<td><em>Raven Relational Scoring By Item.</em> Items are identified by their standard numbers in Raven’s Advanced Progressive Matrices. The Relations column lists which rules are present in the problem matrix and the feature to which that relational rule is applied in parentheses. The last eight columns represent the eight possible responses (moving left to right, top row is 1 2 3 4 and bottom row is 5 6 7 8) and whether or not they contain the corresponding relation.</td>
</tr>
</tbody>
</table>
# List of Figures

<table>
<thead>
<tr>
<th>FIGURE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Example of a Raven problem (left) and a corresponding eye movement scanpath (right). The scanpath is composed of places where the individual looked called fixations (visually depicted as small red circles) and rapid transitions between fixations called saccades (visually depicted as red lines connecting fixations).</td>
</tr>
<tr>
<td>1.2</td>
<td>An example of the type of regularities that can be extracted by the SRSA algorithm (left, the 3x3 red diagonal boxes indicate the benefit of systematically scanning row by row). Chapter 4 actively manipulated how observers view the problem information to quantify the role of strategy (middle). Chapter 4 also tested how relational extraction is affected by actively constraining attentional scope during Raven solution(right).</td>
</tr>
<tr>
<td>2.1</td>
<td>Example of the Raven problem format and trial sequence. Left: The problem matrix and the 8 response alternatives are shown with solid lines. The height of the rectangular box around the matrix subtended 9 degrees of visual angle. Eye fixations were assigned to 10 areas of interest (AOIs) as indicated by dotted lines: nine for the matrix cells (top row = 1–3, middle = 4–6, bottom = 7–9) and one for the entire response area. Right: Each trial had three phases: fixation, solution, and response. Participants fixated for 1 s. Eye movements and verbal protocols were collected during the solution phase. Moving the mouse cursor out of the fixation box triggered the response phase, during which the problem matrix was masked and the participant clicked on their chosen answer. The inter-trial interval (ITI) was 200 ms. (This problem was generated by the authors to protect the security of the standardized test.)</td>
</tr>
</tbody>
</table>
2.2 Principal components, weight matrix, and Raven score prediction for the optimal model described in the text. Component \( a \) captures the tendency to scan the problem matrix row by row (as indicated by the \( 3 \times 3 \) clusters of positive values along the diagonal), whereas component \( b \) penalizes the tendency to toggle to the response area (as indicated by the negative values in the last row). The prediction weight matrix \( (c) \) is the sum of these two components, scaled by their respective regression coefficients. The x-axis represents the sender area of interest (AOI) and the y-axis represents the receiver AOI. Panel \( d \) plots the predicted versus observed Raven scores for all 35 participants \( (R^2=.56) \).

2.3 Synthetic data illustrating the systematic and toggling strategies and their respective successor representations (SRs). Sample fixation sequences generated according to the systematic (\( a \)) and toggling strategies (\( b \)). Panels \( c \) and \( d \) plot the corresponding SR matrices, each averaged across 28 replications. The diagonal box structure in \( c \) reflects the row-by-row scanning pattern in \( a \), whereas the bottom-heavy matrix in \( d \) reflects the toggles to the response area. Matrix \( e \) is the mean of \( c \) and \( d \). Panels \( f \) and \( g \) plot the deviations from the mean—hence the negative (blue) values. Compare with Figure 2.2.

2.4 Leave-one-out cross validation results. The average weight matrix \( (a) \) across 35 leave-one-out fits is virtually identical to the weight matrix produced by the fit to all data at once (Figure 2.2c). Each Raven score was predicted by a separate model that had no access to the data for the respective individual. The squared correlation between the cross-validated predictions and the observed scores was \( R^2_{cv} = .41 \).
3.1 Example of the Raven’s problem format, relational coding, and trial sequence. Left: The problem matrix and the 8 response alternatives are shown with solid lines. The height of the rectangular box around the matrix subtended 9 degrees of visual angle. Eye fixations were assigned to 10 areas of interest (AOIs): nine for the matrix cells (top row = 1–3, middle = 4–6, bottom = 7–9) and one for the entire response area. This example item (generated by the authors) requires the extraction of three relations: distribution of three shapes (diamond, triangle, parallelogram), distribution of three line orientations (0°, 45°, 90°), and quantitative pairwise progression of line numbers (3→2→1). The vectors above each response were not shown to participants but illustrate the respective relations captured in each possible response. Right: Each trial had three phases: fixation, solution, and response. Eye movements and verbal protocols were collected during the solution phase. Moving the mouse cursor out of the fixation box triggered the response phase, during which the problem matrix was masked and the participant clicked on their chosen answer. The inter-trial interval (ITI) was 200 ms.

3.2 Weight matrices, relational score gain predictions for the full cross-validated model, and Simulated SR differences. The cross-validated model prediction weight matrix across 35 leave-one-out fits (a) revealed a strong relationship between systematic scanning and relational score gains across sessions. The relational score gain was predicted by a separate model that had no access to the data for the respective individual. Panel b plots the predicted versus observed relational score gain for all 35 participants ($R^2_{cv} = .32$). Panel c and d were generated using simulated scanpath sequences to highlight important structure. Panel c shows an idealized difference SR resulting from simulated sequences with a 90% increase in row-systematicity on session 2. Panel d shows an idealized difference SR resulting from simulated sequences with a 20% boost in answer checking on session 2. The x- and y-axes represent the sender and receiver areas of interest, respectively. R = response area of interest. SR = successor representation of the regularities in scanpath sequences.
3.3 Comparison of cross-validated prediction weights for high- and low-improvement groups. Relational score gains were predicted separately for the eleven highest and eleven lowest improvement participants across 11 leave-one-out fits (low $R^2_{cv} = .30$; high $R^2_{cv} = .44$). Each value was predicted by a separate model that had no access to the data for the corresponding individual. Panel (a) shows the average prediction weight matrix for the low-improvement group and panel (b) for the high-improvement group. A comparison of the prediction weight matrices shows markedly more diffuse scanning in the low-improvement group (panel a) and a gain in systematicity in the high-improvement group (panel b).

4.1 Example of the Raven problem format and active manipulation paradigms
Left: The problem matrix and the 8 response alternatives are shown with solid lines. The height of the rectangular box around the matrix subtended 9 degrees of visual angle. This problem was generated by the authors to protect the security of the standardized test. Center: Example of the session 2 viewing conditions in Experiment 1. The problem and response areas were outlined to allow easy orientation within each item. Only one cell of the problem could be viewed at any given time and the order of problem information was either row-wise systematic, haphazard, or free. The blue lines indicate possible systematic transitions given the current cell while the red arrows indicate possible haphazard transitions. In the control condition the participant could freely transition to any cell using the mouse. Right: Example of the session 2 viewing conditions in Experiment 2. Participants viewed each item through a gaze-contingent viewing window (diameter subtended 3.2 degrees of visual angle) that allowed only one cell of the problem to be viewed at any given time.

4.2 RAPM score and solution time during pretest and active strategy manipulation by condition
Left: The number of APM items answered correctly during pretest and for free-view, haphazard, and systematic scanning conditions on session 2. The constrained free-view was not significantly different than pretest. Haphazard scanning produced a significant deficit in APM score and systematic row-wise scanning produced a significant increase in APM score. Error bars represent standard error of the mean. Right: Total APM solution time during pretest and for free-view, haphazard, and systematic scanning conditions on session 2. Significant increases in solution time were observed for the haphazard and systematic viewing conditions relative to pretest. Error bars represent the standard error of the mean.
4.3 APM score and solution time during pretest and gaze-contingent test sessions

Left: The number of APM items answered correctly during pretest and gaze-contingent (GC) test sessions across all participants. No significant differences between pretest and gaze-contingent performance were observed. Error bars represent the standard error of the mean. Right: Total APM solution time during pretest and gaze-contingent (GC) viewing across all participants. No significant effect of gaze-contingent viewing was found for solution time, suggesting the increases in solution time observed in Experiment 1 were likely driven by the keyboard and mouse interfacing rather than the reduction in problem information actively available. Error bars represent the standard error of the mean.

5.1 Photograph of the artificial eye, ocular socket, and pupil calibration devices. Each artificial eye was measured in a constant left-eye position within the chin- and forehead-rest. The pupil calibration apparatus was not present during artificial-eye data collection as shown, but rather replaced the ocular socket during the pupil calibration procedure.

5.2 Schematic illustrations of the calibration apparatus, artificial eye sizes, laser orientation apparatus, ocular socket, and target displays. The windowed pupil calibration apparatus (a) allowed steady presentation of each of 7 black dots of known diameter printed on a paper slip. (b) Three spherical artificial eyes were constructed with pupil diameters 3.17 mm, 4.76 mm, and 7.14 mm, respectively. The gaze direction of the artificial eyes were precisely guided using an attached laser pointer (c) whose collimated beam illuminated a bright dot on the target display screen (e). A simple socket (d) allowed the artificial eye to rotate to “fixate” a target while maintaining constant physical position.

5.3 Experimental Layouts and Geometric Model. Three separate geometric layouts (panels a–c) varied the relative positions of the camera, artificial eye, and monitor. Each layout is diagrammed from three vantage points (top, side, and front). All distances are in millimeters. Pupil calibration (Experiment 1) and pupil foreshortening (Experiment 2) were measured for each layout. The geometric model (d) estimates the foreshortening of the pupil area as a function of the cosine of the angle \( \theta \) between the eye-to-camera axis \( OC \) and the eye-to-target axis \( OT \). The origin \( O \) of the coordinate system is at the center of the artificial pupil. The \( X \) axis is horizontal and parallel to the bottom edge of the screen, growing rightward. The \( Y \) axis is vertical, growing upward. The \( Z \) axis is perpendicular to the screen, growing inward.
5.4 *Ratio scale between arbitrary units and physical units* The pupil-diameter data recorded by EyeLink 1000 in arbitrary units are proportional to the true diameter of the physical pupil, but the coefficient of proportionality depends on the distance between the camera and the artificial eye.

5.5 *Deviation in pupil diameter by pupil size and layout.* Each panel shows the deviation in measured pupil diameter from the geometric mean for a unique combination of pupil size (3.17 mm, 4.76 mm, 7.14 mm) and experimental layout (Near, Medium, Far). Large deviations in measured pupil diameter were observed across each map despite the fixed pupil of each artificial eye. Deviations in measured pupil size grew as layout distance became shorter and visual angles increased. Deviations were relatively invariant across pupil diameter within a given experimental layout. Gaze positions are in degrees of visual angle.

5.6 *Aggregate empirical maps, parameter-free geometric model multipliers, and parameter-free correction* Each panel shows the deviation in pupil diameter from the geometric mean. The first column of panels shows the aggregate empirical maps collapsed across pupil size for each experimental layout. The second column shows the corrective multipliers produced by the parameter-free geometric model. The third column shows the corrected data produced by dividing the aggregate empirical data by its respective corrective multiplier. The parameter-free geometric correction reduced the within-map RMSE to 18.7%, 15.8%, and 18.2% of its original value for the Near, Medium, and Far layout. Gaze positions are in degrees of visual angle.

5.7 *Aggregate empirical maps, optimized geometric model multipliers, and optimized correction* Each panel shows the deviation in pupil diameter from the geometric mean. The first column of panels shows the aggregate empirical maps collapsed across pupil size for each experimental layout. The second column shows the corrective multipliers produced the geometric model with optimized parameters. The third column shows the corrected data produced by dividing the aggregate empirical data by its respective corrective multiplier. The optimized geometric correction virtually eliminated pupil foreshortening error, reducing the within-map RMSE to 2.0%, 2.5%, and 3.2% of its original value for the Near, Medium, and Far layout. Gaze positions are in degrees of visual angle.
6.1  **Raven problem format and trial sequence.** Left: The problem matrix and the 8 response alternatives are shown with solid lines. The height of the rectangular box around the matrix subtended 9 degrees of visual angle. This example item (generated by the authors) contains three relations that must be extracted: distribution of three shapes (diamond, triangle, parallelogram), distribution of three line orientations ($0^\circ$, $45^\circ$, $90^\circ$), and decreasing line number down columns ($3\to2\to1$). Right: Each trial had three phases: fixation, solution, and response. Participants fixated for 1 s. Eye movements and concurrent think aloud verbal protocols were collected during the solution phase. Moving the mouse cursor out of the fixation box triggered the response phase, during which the problem matrix was masked and the participant clicked on their chosen answer. The inter-trial interval (ITI) was 200 ms.

6.2  **Mean change in pupil diameter by fluid intelligence and**  Left: Comparisons of the mean percent change in pupil diameter from baseline for exploratory and exploitative periods is shown averaging across all participants and for subgroups of participants with APM scores of less than 8, 8-10, 11-12, and 13-14 ($N=5$ for each subgroup). The error bars represent the standard error of the mean. Right: Scatterplot of APM score and mean % change in pupil diameter during exploration. A linear regression revealed that mean exploratory pupil diameter explained 25% of the variance in APM score.
How are humans able to reason so effectively when confronted with novel, relational environments? The long-term objective is to gain insight into this question and the underlying mechanisms of visual relational reasoning by studying how people solve visual analogy problems from a benchmark test of fluid intelligence, Raven’s Advanced Progressive Matrices (APM) (Raven, Raven, & Court, 1998). The Raven APM is a geometric analogy test with excellent psychometric properties that has, for 70 years, been a popular and trusted instrument across the field of psychology (Gray, Chabris, & Braver, 2003; Brouwers, Van de Viver, & Van Hemert, 2009; Eslinger et al., 2009). Performance on fluid intelligence tests like the Raven APM are of great functional importance because fluid ability is one of the best predictors of job performance and educational achievement (Gottfredson, 1997) and represents an important dimension of individual differences (Cattell, 1971). Each Raven item consists of a 3x3 problem matrix which contains relations across each row and down each column of the matrix and the goal is to solve for the missing cell (see Figure 1.1, left). There are 8 possible response alternatives at the bottom of the image, one of which is the best answer. Raven performance and fluid ability are thought to be closely related to attentional control (Kane & Engle, 2002; Engle, Kane, & Tuholski, 1999) and working memory (Conway, Cowan, Bunting, Therriault, & Minkoff, 2002; Engle, Tuholski, Laughlin, & Conway, 1999). Here individual differences in a rich set of data including relational
ability (i.e. the number of items correct/relations extracted), solution time (i.e. the
time it took to solve each Raven item), eye movement patterns (i.e. where individuals
looked on the image during solution), verbal protocols (i.e. individuals verbally ex-
plaining how they are solving each item), and pupil diameter (used as a non-invasive
index of noradrenergic activity) were used to study attentional control and working
memory processes during relational reasoning. The main objectives of the current
research were quantifying the role of information processing strategy during rela-
tional reasoning, assessing the contribution of strategy refinement in the test-retest
designs commonly used to study cognitive enhancement, establishing a causal rela-
tionship between strategic processing and the probability of relational insight, and
understanding how the mediation of the exploration-exploitation tradeoff contributes
to relational reasoning ability.

Figure 1.1 Example of a Raven problem (left) and a corresponding eye movement scanpath
(right). The scanpath is composed of places where the individual looked called fixations
(visually depicted as small red circles) and rapid transitions between fixations called saccades
(visually depicted as red lines connecting fixations).
Eye movements served as an important data source for studying attentional control during relational reasoning because they provide a rich description of the strategic allocation of attention over time (Findlay & Gilchrist, 2003). One major obstacle to studying information processing strategy in visual relational environments is developing a methodology that is capable of extracting statistical regularities from highly complex sequences of eye movements (called scanpaths, see Figure 1.1, right). Scanpaths are difficult to analyze because the number of possible scanpaths grows exponentially with their length. To illustrate, suppose the display is divided into 10 areas of interest (AOIs). Then there are 10 scanpaths of length 1 but 590,490 (= \(10 \times 9^5\)) scanpaths of length 6. The challenge is to tame this combinatorial explosion without losing the sequential information in the process. The pre-existing methods for analyzing scanpaths have significant shortcomings, which limited their usefulness in studying relational reasoning (Ponsoda et al., 1995; Simola et al., 2008). Therefore, in Chapter 2 a novel methodology for analyzing scanpaths is developed and deployed that is capable of extracting interpretable regularities from the complex patterns of eye movements that are typically seen in visual relational reasoning contexts like Raven (called Successor Representation Scanpath Analysis or SRSA, Dayan, 1993; Sutton & Barto, 1998; Hayes, Petrov, & Sederberg, 2011). The SRSA technique was able to successfully extract interpretable strategies from sequential eye movement patterns and predict individual differences in Raven score with unprecedented precision—providing important new insights into the role of processing strategy and attentional control during relational reasoning.

Pretest-train-posttest designs using the Raven APM or similar matrix-based reasoning tests have become standardly used tools to assess gains in fluid intelligence in the rapidly growing field of cognitive enhancement (Morrison & Chein, 2011; Melby-Lervåg & Hulme, 2013). For the most part, attempts to alter fluid intelligence through
training have yielded disappointing results for healthy adults (Detterman & Sternberg, 1982). However, recent reports of training-induced gains on Raven following working memory training (Jaeggi, Buschkuehl, Jonides, & Perrig, 2008) have fueled an explosion of interest in cognitive training — now a billion-dollar industry (“Brain sells,” 2013). The study by Jaeggi et al. (2008) was one of the first to show evidence that fluid intelligence could be improved with training—suggesting that fluid intelligence may not be the stable, primarily heritable trait previously thought. The study used a pretest-train-posttest design with an untrained control group. A titrated, adaptive dual n-back task was practiced for up to 18 sessions in the experimental group (\(N = 34\)) but not in the control group (\(N = 35\)). All participants were pre- and post-tested on two parallel short-form versions of a matrix-based \(Gf\) test—either Raven’s Advanced Progressive Matrices (Raven et al., 1998) or BOMAT (Hossiep, Turck, & Hasella, 1999). Whereas the results showed statistically significant score gains in both groups, the average gain in the trained group was significantly higher than that in the control (Jaeggi et al., 2008). The latter finding—a significant control-adjusted gain—was interpreted as an improvement in \(Gf\) and fueled the current boom in the cognitive enhancement industry, as well as a big controversy in the scientific literature. Unfortunately, subsequent attempts to replicate the putative improvement in \(Gf\) have produced mixed results (e.g., Chooi & Thompson, 2012; Harrison et al., 2013; Jaeggi, Buschkuehl, Jonides, & Shah, 2011; Jaeggi et al., 2010; Redick et al., 2012; Thompson et al., 2013) and it is unclear why. This rapidly growing field is characterized by large variations in reported effect sizes (see Melby-Lervåg & Hulme, 2013, for a meta-analysis of 23 studies), polarization of opinion, and contradictory reviews (e.g., Buschkuehl & Jaeggi, 2010; Morrison & Chein, 2011, on the optimistic side; Melby-Lervåg & Hulme, 2013; Shipstead, Redick, & Engle, 2012, on the skeptical side). Often the focus in trying to understand replications and failures to replicate
focus on the minutia of the training regimens or aspects of the control groups (e.g. active vs. passive controls to account for motivational effects), while little attention has been paid to the difficulties in assessing the change in latent variables like fluid intelligence using test-retest designs.

In Chapter 3, SRSA is used to gain important new insights into this critical applied domain. Specifically, SRSA is used to examine the role of strategy refinement in test-retest improvement in Raven scores to determine whether we really become smarter when our fluid-intelligence test scores improve? The SRSA analysis from 35 participants solving Raven items on two separate sessions indicated that one-third of the variance of score gains could be attributed to refinements in test-taking strategy alone, as revealed by characteristic changes in eye-fixation patterns. Participants that improved on retest scanned problem matrix more systematically and checked their answer more carefully, while participant that got worse scanned the problem matrix more haphazardly. When the strategic contaminant was partialled out, the residual Raven score gains were no longer significant. These results while not definitive, raise the possibility that the score-gains on matrix-reasoning tests typically used to measure the efficacy of cognitive training regimens may reflect strategy refinement instead of true intelligence gains—particularly when visual training regimens are used (Moody, 2009; Stephenson & Halpern, 2013; Jaeggi, Buschkuehl, Shah, & Jonides, 2014).

The results from Chapter 2 indicated that broad strategies can be extracted from eye movements and used to predict how many Raven problems participants will get correct (Hayes et al., 2011). Specifically, the findings showed a benefit to systematically scanning the problem information and a detriment to frequently switching between problem and response information. Chapter 3 revealed that those that improve the most in test-retest designs scan the problem matrix more systematically at posttest and those that get worse scan the problem matrix more haphazardly. There
are multiple possibilities for why systematic scanning is correlated with higher scores. One explanation is that higher ability participants are simply more likely to systematically scan the problem than lower ability participants. Another possibility is that systematically scanning each row actually increases the probability that a participant will successfully extract the correct relation or relations. Previous studies in other types of insight problem solving have suggested that the way a person visually scans a problem may increase the likelihood of rule insight (Grant & Spivey, 2003; Thomas & Lleras, 2007). To address this issue, experiment 1 in Chapter 4 actively manipulated the level of systematicity of element viewing to establish a causal relationship between strategy and the probability of relational insight—quantifying the benefits and deficits caused by systematic and haphazard scanning. Problem information was presented one cell at a time either systematically by row or randomly (see Figure 1.2, middle panel).

Actively manipulating the order in which Raven problem information was presented to the observer required a reduction in the information available to visual attention and working memory—an interesting manipulation in its own right. To understand the effects of this manipulation in isolation, a separate, gaze-contingent paradigm was also employed that altered the displayed image in real-time based on where the observer is looking. This limited viewing manipulation presented in Experiment 2 of Chapter 4 places additional demands on working memory and constrains visual attention allowing insights into the relationships between working memory capacity, attentional scope, and fluid intelligence (Fry & Hale, 2000; Gray & Thompson, 2004; Halford, Cowan, & Andrews, 2007; Kane & Engle, 2002).

The ability to adaptively regulate the balance between exploration and exploitation is critical for optimizing performance in novel, relational environments like Raven. Within each novel relational environment participants must both explore and exploit,
Figure 1.2 An example of the type of regularities that can be extracted by the SRSA algorithm (left, the 3x3 red diagonal boxes indicate the benefit of systematically scanning row by row). Chapter 4 actively manipulated how observers view the problem information to quantify the role of strategy (middle). Chapter 4 also tested how relational extraction is affected by actively constraining attentional scope during Raven solution (right).

as they shift between testing specific relational hypotheses and exploring other options when a given hypothesis fails to lead to a possible item solution. Converging evidence from primate electrophysiology (Aston-Jones et al., 1991, 1997, 1994) and computational neural modeling (Usher, Cohen, Servan-Schreiber, Rajkowski, & Aston-Jones, 1999; Servan-Schreiber, Printz, & Cohen, 1990) have indicated that changes in exploratory versus exploitive control state may be mediated by the broad noradrenergic projections emanating from the locus coeruleus (LC). The LC is a subcortical nucleus in the dorsorostral pons that serves as the principal source of NE for a wide range of brain areas including the amygdala, thalamus, hippocampus, and neocortex (Bouret & Sara, 2005; Aston-Jones & Cohen, 2005). The synthesis of these findings has resulted in the development of a theory of LC function (adaptive gain theory, Aston-Jones & Cohen, 2005) in which the LC modulates the gain of cortical units to facilitate either exploration or exploitation in response to current assessments of task utility.
At the same time, pupil diameter has recently emerged as a promising noninvasive proxy measure for LC activity, and is increasingly being employed for this purpose (e.g., Cheadle et al., 2014; Einhäuser, Koch, & Carter, 2010; Wolfgang, Stout, Koch, & Carter, 2008; Eldar, Cohen, & Niv, 2013; Gilzenrat, Nieuwenhuis, Jepma, & Cohen, 2010; Jepma & Nieuwenhuis, 2011). Converging evidence from electrophysiology (e.g., Rajkowski, Kubiak, & Aston-Jones, 1994), pharmacology (e.g., Koss, 1986; Phillips, Szabadi, & Bradshaw, 2000), and neuroimaging (Murphy, O’Connell, O’Sullivan, Robertson, & Balsters, 2014) suggest that pupil diameter under constant illumination is correlated with LC activation in animals and humans. The anatomical pathways linking the locus coeruleus and the pupil are a topic of ongoing research, but probably involve α2-adrenoreceptor mediated inhibition of the parasympathetic Edinger-Westphal nucleus responsible for pupil constriction (Samuels & Szabadi, 2008a, 2008b).

While Raven is well-suited for studying exploration-exploitation tradeoffs, it also posed a major obstacle to measuring pupillary response: pupil foreshortening error (PFE). PFE is caused by changes in gaze position which produce foreshortening of the pupillary image because the eye tracking camera is fixed but the eye rotates. As the eye rotates away from the camera, the pupil image becomes more elliptical and shrinks in apparent area by as much as 10%—a large margin given that most cognitive pupillary effects produce less than a 5% change in pupillary response. PFE is a potentially serious confound that must be controlled because Raven solution requires participants freely view each Raven item. Existing methods for correcting PFE across changes in gaze position only reduced PFE by about 10% (Brisson et al., 2013).

Therefore, in order to use pupillometry as an indirect index of LC activity, PFE was systematically mapped and a new correction method was developed that virtually
eliminated PFE. Three artificial eyes with different fixed pupil sizes were used to systematically measure changes in pupil size as a function of gaze position. A grid-based map of pupil measurements was recorded with each artificial eye across 3 experimental layouts of the eye tracking camera and display. Large, systematic deviations in pupil size were observed across all 9 maps. The measured PFE was corrected by a geometric model that expressed the foreshortening of the pupil area as a function of the cosine of the angle between the eye-to-camera axis and the eye-to-stimulus axis. The model reduced the root mean squared error of pupil measurements by 82.5% when the model parameters were pre-set to the physical layout dimensions, and by 97.5% when they were optimized to fit the empirical error surface. The optimized fit which reduced PFE by 97.5% allowed the measurement of cognitive pupillary effects in Chapter 6.

Chapter 6 pioneers the use of the pupillary response in conjunction with verbal protocol analysis to understand how the mediation of the exploration-exploitation tradeoff contributes to visual relational reasoning. Think aloud verbal protocols (Ericsson & Simon, 1993), in which participants explain in real-time how they are solving each Raven item were used to identify exploratory and exploitative solution periods. Pupil diameter was used as a non-invasive index of noradrenergic activity and was compared across exploratory and exploitative periods across individual differences in fluid intelligence. The results showed an increase in pupil diameter during exploratory periods and decrease during exploitative periods consistent with prominent theories of LC function (Aston-Jones & Cohen, 2005). In addition, individual differences in fluid intelligence were linearly related to the increase in pupillary response during exploratory periods. To our knowledge this is the first study showing evidence that noradrenergic activity may mediate the exploration-exploitation tradeoff
during analogical reasoning and the only study to show an exploration-related boost in pupillary response that is modulated by individual differences in fluid ability.
A novel method for analyzing sequential eye movements reveals strategic influence on Ravens Advanced Progressive Matrices

2.1 Introduction

Eye-movement protocols are an important data source in vision science and psychology (e.g., Buswell, 1935; Yarbus, 1967) and have advanced our knowledge of visual search, scene perception, development, human-computer interaction, reading, and many other fields (see, e.g., Findlay & Gilchrist, 2003; Rayner, 1998, for reviews). Despite this success, the vast majority of eye-movement studies have ignored all sequential information in the data and utilized only first-order statistics such as fixation probabilities and dwell times. Although fixation sequences (or scanpaths, Stark & Ellis, 1981) often contain valuable information about underlying cognitive processes, they are difficult to quantify and interpret, and this has traditionally prevented eye-tracking researchers from including them in their analyses.

Why are scanpaths so difficult to analyze? The fundamental reason is that the number of possible scanpaths grows exponentially with their length. To illustrate, suppose the display is divided into 10 areas of interest (AOIs). Then there are 10

The text of this chapter [2] is taken from Hayes, Petrov, & Sederberg, 2011

11
scanpaths of length 1 but 1,000,000 scanpaths of length 6. The challenge is to tame this combinatorial explosion without losing the sequential information in the process. The existing methods for doing this can be classified into two broad classes. One approach represents scanpaths as strings of letters and uses string-editing distance—the number of additions and subtractions necessary to turn one sequence of letters into another—as a dissimilarity metric (e.g., Brandt & Stark, 1997; Myers & Schoelles, 2005). Yet, string-editing measures have a number of limitations, with the most critical being that they are best-suited for comparing short sequences of similar length, making it difficult to infer cognitive states or strategies in temporally extended tasks or to compare across participants or trials that differ in duration.

Another approach is based on transition probability matrices (e.g., Ponsoda et al., 1995; Ellis & Stark, 1986) and Markov models (e.g., Jansen, Marriott, & Yelland, 2007; Salvucci & Anderson, 2001; Simola et al., 2008), which can be used to extract and compare regularities in scanpaths of varying length. This method also has limitations. While a transition matrix provides a relatively simple representation of scanpath information (a fixed-size matrix), it only estimates the conditional probabilities of scanpaths of length 2. That is, given a current fixation on one AOI, what is the probability to visit each of the other AOIs on the next fixation? This is a very limited event horizon—reaching only one step into the future. Higher-order transition matrices extend the horizon to two steps (or more) but there seldom are enough data to provide accurate estimates of the (exponentially growing number of) higher-order probabilities. Hidden Markov models (HMMs, e.g., Rabiner, 1989) deal with the combinatorial explosion by factoring the joint probability density into smaller, more manageable pieces using conditional-independence assumptions. When these assumptions are met, HMMs have been applied successfully in eye-movement
data analysis (e.g., Cagli, Coraggio, Napoletano, & Boccignone, 2008; Salvucci & Anderson, 2001; Simola et al., 2008; van der Lans, Pieters, & Wedel, 2008) and active computer vision (e.g., Rimey & Brown, 1991). The factorization is formalized in a graphical model whose parameters are then estimated from data via sophisticated algorithms such as Markov chain Monte Carlo (e.g., Scott, 2002; van der Lans et al., 2008). This makes the development of a HMM a slow and laborious process that requires domain knowledge and considerable expertise. This method seems ill suited for exploratory data analysis in domains where the underlying factorization is not known.

The present article pioneers the use of reinforcement-learning algorithms to capture temporally-extended sequential information in eye-movement protocols. We present a novel application of a temporal-difference learning algorithm (Sutton, 1988; Sutton & Barto, 1998) to construct a successor representation (SR, Dayan, 1993; White, 1995) of an eye-movement sequence that keeps the simplicity of the fixed-size transition matrix and extends the event horizon. The key idea is that upon observing a transition from one AOI to another, instead of simply updating the transition probability from the first to the second AOI, we associate the first AOI with the second AOI and all expected subsequent AOIs based on prior visits to the second AOI. This is equivalent to learning to predict future scanpaths based on past scanpaths. After traversing the entire fixation sequence for a trial, the resulting SR can be conceptualized as having extracted the statistical regularities in temporally-extended scanpaths, collapsing the information into a fixed-size matrix. Specifically, an SR matrix contains, for each AOI, the temporally discounted number of expected future fixations to all AOIs (Dayan, 1993). Given their uniform size the SR matrices from different observers and/or trials can be analyzed using standard statistical methods to identify
significant regularities for various comparisons of interest. The new method is very well suited for exploratory data analysis.

To demonstrate the effectiveness of the scanpath SR as an exploratory tool, we apply this method to discern individual differences in problem-solving strategies on a benchmark test of fluid intellectual ability, Raven’s Advanced Progressive Matrices (APM, Raven et al., 1998). The Raven APM is a geometric analogy test with excellent psychometric properties (Brouwers et al., 2009) that has, for 70 years, been a popular and trusted instrument in clinical (e.g., Soulieres et al., 2009), developmental (e.g., Eslinger et al., 2009), and cognitive (e.g., Gray et al., 2003) psychology. As we report in the Results below, the SR analysis allows us to predict individual Raven scores with unprecedented precision from the eye-movement data. In the process, the scanpath SR also yields new theoretical insights about Raven problem-solving strategies.

We can evaluate scanning patterns to predict Raven scores because both measures correlate with a third, hidden variable—strategy. Individuals differ in their problem-solving strategies and this is detectable in their eye movements (e.g., Just & Carpenter, 1985). A Raven problem consists of a matrix and 8 response alternatives (Figure 2.1, left). Two strategies are particularly relevant for such problems (Snow, 1980). In constructive matching, the participant tries to formulate the missing element based exclusively on matrix information, and then looks for that element in the response area. In response elimination, each alternative is inspected in turn and evaluated whether it fits into the empty matrix slot. The former strategy tends to occur in high-scoring individuals and/or easier problems, the latter in low-scoring individuals and/or difficult problems (Bethell-Fox, Lohman, & Snow, 1984; Vigneau et al., 2006). We will show that the scanpath SR identifies the degree to which participants apply each of these two strategies. This can then be used to predict the individual scores on the Raven task.
Figure 2.1 *Example of the Raven problem format and trial sequence.* Left: The problem matrix and the 8 response alternatives are shown with solid lines. The height of the rectangular box around the matrix subtended 9 degrees of visual angle. Eye fixations were assigned to 10 areas of interest (AOIs) as indicated by dotted lines: nine for the matrix cells (top row = 1–3, middle = 4–6, bottom = 7–9) and one for the entire response area. Right: Each trial had three phases: fixation, solution, and response. Participants fixated for 1 s. Eye movements and verbal protocols were collected during the solution phase. Moving the mouse cursor out of the fixation box triggered the response phase, during which the problem matrix was masked and the participant clicked on their chosen answer. The inter-trial interval (ITI) was 200 ms. (This problem was generated by the authors to protect the security of the standardized test.)
2.2 Experiment

2.2.1 Method

Thirty-five university students with normal or corrected-to-normal vision completed 28 problems from Raven’s Advanced Progressive Matrices, Set II (Raven et al., 1998) on two separate days approximately a week apart. The participants were paid $6 per hour plus $1 bonus for each correct answer. Half of them completed items 2, 4, 6, 9, 10, 11, 16, 17, 19, 21, 23, 24, 26, and 29 on the first session and 1, 3, 5, 7, 12, 13, 14, 15, 18, 20, 22, 25, 27, and 28 on the second. The other half completed the same subsets in the opposite order. The instructions followed the Raven APM Manual guidelines for individual test administration (Raven et al., 1998).

A chin rest was located $\approx 92$ cm away from the 21” CRT monitor in a darkened room. Each trial began with a brief alert sound and a fixation cross appeared in the middle of the screen (Figure 2.1, right). After the participant fixated for 1 s, which allowed for equipment re-calibration, the Raven problem appeared and the participant had unlimited time to work on it. A mouse click on one of the responses ended the trial.

Eye-tracking data and “think aloud” protocols (Ericsson & Simon, 1993) were collected on both sessions. Between these main sessions, 23 of the participants completed two additional sessions of paper-and-pencil practice on Raven-like problems (Matzen et al., 2010). This manipulation had no statistically significant effect relative to a control group\(^1\) ($F(2,32) = 0.15$, $p = .86$). Therefore, we analyzed the test data of

\(^1\)There was a significant practice effect within subjects but it did not interact significantly with the between-subject manipulation. The posttest score was 1.5 points higher on average than the pretest score ($t(34) = 3.48$, $p < .001$), replicating published results (Denney & Heidrich, 1990; Bors & Vigneau, 2003).
all 35 participants together. The paper-and-pencil data and the verbal protocols are beyond the scope of this article.

**Sequential Eye Movement Analysis**

Participants’ eye movements were recorded using an Eyelink 1000 desktop eye tracker (SR Research, 2006) at a sampling rate of 1000 Hz. Saccades and fixations were segmented with Eyelink’s standard algorithm using velocity and acceleration thresholds (SR Research, 2006). Each fixation was assigned to one of the 10 AOIs depicted in Figure 2.1. The few (<1%) fixations that fell outside of the 10 designated areas were ignored. A single AOI (labeled R) covered the entire response area so that the spatial layout of the answers could not be used to predict the participants’ scores.

We defined a scanpath as the sequence of fixations across the 10 different AOIs on a given trial. The sequences varied widely in length across participants and trials. In an effort to reduce this variability, we clipped 20% from the beginning of each sequence longer than 100 fixations. If the clipped length still exceeded 100, we also clipped 5% from the end. The median length of the clipped scanpaths used in the analyses was 88 fixations (min=14, max≈1000, IQR=69). The clipping also helped to isolate the period of active problem solving, given that the early fixations tended to survey the matrix and the last few verified the chosen answer.

The next step was to calculate the successor representation (SR; Dayan, 1993) for each scanpath. We used a temporal-difference learning algorithm to extract long-range statistical regularities from the sequence. The algorithm treats each scanpath as a first-order Markov chain with the 10 AOIs comprising a discrete, finite state space (Dayan, 1993; White, 1995). The algorithm is incremental and builds a $10 \times 10$ SR matrix $M$. The matrix is initialized with zeros and then updated for each transition
in the sequence. Consider a transition from state \( i \) to state \( j \). The \( i \)th column of the matrix—the column corresponding to the “sender” AOI—is updated according to:

\[
\Delta M_i = \alpha(I_j + \gamma M_j - M_i),
\]

where \( I \) is the identity matrix, each subscript picks a column in a matrix, \( \alpha \) is a learning-rate parameter \((0 < \alpha < 1)\), and \( \gamma \) is a temporal discount factor \((0 < \gamma < 1)\).

In words, upon observing a transition \( i \rightarrow j \), the set of expected successors \( (M_i) \) for the sender \( i \) is updated to include the receiver \( j \) (represented as a unit column-vector \( I_j \)) and the predicted set of successors \( (M_j) \) for the new location \( j \), discounted by \( \gamma \). The latter term is the key to extending the event horizon to encompass both immediate and long-range transitions—it includes the discounted future states in the prediction from the current state. For example, suppose a participant scans the top row of a Raven problem systematically from left to right: \( 1 \rightarrow 2 \rightarrow 3 \rightarrow 1 \rightarrow 2 \ldots \) Then the successors of location 1 will include both location 2 and, weighted by \( \gamma \), location 3. By contrast, a first-order transition matrix would include only the association between 1 and 2. After traversing the whole scanpath, the estimated SR matrix approximates the ideal SR matrix, which contains the temporally discounted number of expected future fixations on all AOIs (rows), given the participant just fixated on any individual AOI (column). Note that the entries in the SR matrix are not probabilities; They are (discounted, expected) \textit{numbers of visits} and thus the sum across each column of the ideal SR matrix equals:

\[
1 + \gamma + \gamma^2 + \ldots = \frac{1}{1-\gamma} \geq 1
\]

Appendix A provides additional technical details.

To summarize, given parameters \( \alpha \) and \( \gamma \), the algorithm produced one \( 10 \times 10 \) SR matrix per participant per trial. Averaging across the 28 trials for each participant, we were left with 35 individual matrices. Each matrix summarized the eye-fixation
patterns of the corresponding participant. To reduce the dimensionality of the space, we performed a principal-component analysis (PCA, Everitt & Dunn, 2001) of the successor representations. Each SR matrix was reshaped to a vector of 100 features. The whole data set occupied a matrix of size $35 \times 100$. Following standard PCA practice, we re-scaled each feature (column) so that it had zero mean and unit variance across the 35 participants. The first 20 principal components retained over 90% of the variance in the SR data. Conceptually, these components represent dimensions of individual differences in fixation patterns. They are expressed mathematically as orthogonal basis vectors in the 100-dimensional SR space. Each participant was characterized by 20 projections onto this rotated basis.

Finally, the cumulative Raven score (i.e., the number of correct responses) of each participant was introduced as the target variable of a hierarchical linear regression analysis. The SR projections entered as predictor variables.

We also compared the novel SR method to several regression models with traditional predictors based on AOI dwell times. Following Vigneau et al. (2006), we explored the following variables: proportional time on matrix (PTM = the dwell time on the matrix area divided by the overall latency), latency to first toggle (FT = the time stamp of the first saccade to the response area), overall latency on easy items (LEz), the number of toggles on easy items (NT), the toggle rate on easy items (TR = NT divided by item latency), and matrix time distribution index (MTDI = the proportional dwell time on cells 1, 2, 4, and 6 minus the proportional dwell time on cells 3, 6, 7, 8, and 9). An item was defined as “easy” if at least 80% of the participants answered it correctly (Vigneau et al., 2006). The first 5 items in each of our test sets met this criterion in our data. PTM, FT, and MTDI were averaged across all 28 items, and LEz, NT, and TR across the 10 easy items. Thus, each participant was characterized by 6 measures, which were then used to predict their Raven score.
2.2.2 Results and Discussion

The Raven scores varied between 12 and 27 across the 35 participants (M=21.9, SD=3.7). We performed a hierarchical linear regression to assess how much of this variance can be explained on the basis of the SR principal-component projections. Two of the projections correlated very strongly with the scores, whereas the third-best predictor was insignificant. Therefore, we used two predictors in all regressions. We implemented a two-tier algorithm to maximize the fit to the Raven scores. In the inner loop, it calculated the SR matrices for given parameters $\alpha$ and $\gamma$ (Equation 2.2.1), then calculated the first 20 principal components and the corresponding projections for each participant, picked the two projections that correlated most strongly with the scores, and constructed a linear regression model with these two predictors. In the outer loop, a Nelder-Mead optimization routine searched for $\alpha$ and $\gamma$ that maximized the multiple regression coefficient of the inner-loop model. The best fit ($R^2 = .56$) was achieved with learning rate $\alpha^* = .233$ and discount factor $\gamma^* = .255$. Figure 2.2d reports this optimal model. To our knowledge, this is the most accurate prediction of Raven scores based on eye-tracking data reported to date.

In addition to providing an accurate prediction of Raven scores, the two scanpath SR principal components selected for the regression had clear interpretations with respect to participants’ strategies. Figure 2.2a shows the first component, which accounted for the largest proportion (31%) of the variance in the scores. It was also the first PCA component, capturing the strongest individual differences in eye-movement patterns. This component is characterized by a prominent diagonal “box” structure (Fig. 2.2a). The $3 \times 3$ red boxes indicate the benefits of systematically scanning within a given row of the problem matrix as opposed to haphazard scanning.

---

2The mean scores (and SDs) for the two 14-item subsets were 10.7 (2.8) and 11.2 (1.8). The subsets were counterbalanced across the first (M=10.2, SD=2.5) and second (M=11.7, SD=2.0) sessions.
Figure 2.2 Principal components, weight matrix, and Raven score prediction for the optimal model described in the text. Component a captures the tendency to scan the problem matrix row by row (as indicated by the $3 \times 3$ clusters of positive values along the diagonal), whereas component b penalizes the tendency to toggle to the response area (as indicated by the negative values in the last row). The prediction weight matrix (c) is the sum of these two components, scaled by their respective regression coefficients. The x-axis represents the sender area of interest (AOI) and the y-axis represents the receiver AOI. Panel d plots the predicted versus observed Raven scores for all 35 participants ($R^2=.56$).

or column-wise scanning. The positive (red) values “dripping” from each box indicate systematic integration as participants moved from row to row.

The second component, which accounted for another 25% of the variance in the scores, is dominated by a solid blue line across the response area (Fig. 2.2b). We interpret this solid blue area as an “anti-toggle” component. That is, participants who made fewer toggles from each cell of the problem matrix to the response area achieved higher scores than participants who toggled more frequently.

Figure 2.3 illustrates these two strategies on synthetic data. We generated 28 sequences according to the systematic strategy. Each sequence began with 50 fixations within the 3 AOIs on the first row, followed by 50 fixations within the second row, 50 fixations within the third row, and a few fixations to and from the response
Figure 2.3 Synthetic data illustrating the systematic and toggling strategies and their respective successor representations (SRs). Sample fixation sequences generated according to the systematic (a) and toggling strategies (b). Panels c and d plot the corresponding SR matrices, each averaged across 28 replications. The diagonal box structure in c reflects the row-by-row scanning pattern in a, whereas the bottom-heavy matrix in d reflects the toggles to the response area. Matrix e is the mean of c and d. Panels f and g plot the deviations from the mean—hence the negative (blue) values. Compare with Figure 2.2.
area. Figure 2.3a plots one of those sequences. We calculated 28 SR matrices from these sequences using the optimal parameters $\alpha^*$ and $\gamma^*$. Figure 2.3c plots the average of these matrices. It represents a “pseudo-observer” who consistently follows the systematic strategy on all trials. The diagonal box structure is clearly visible (cf. Fig. 2.2a). Note that the cells along the main diagonal have positive values even though the fixation sequences contained no transitions from any AOI directly back to itself. This illustrates an important difference between the successor representation and a transition probability matrix. Despite the absence of immediate repetitions in the sequence, there are plenty of round-trip scanpaths, which give rise to the positive SR values along the diagonal. We also generated 28 sequences of length 150 according to the toggling strategy. They contained multiple transitions to and from the response area (Fig. 2.3b). The corresponding trial-averaged SR matrix (Fig. 2.3d) has high values along the bottom and right edges, corresponding to scanpaths ending in and starting from R, respectively. Panels f and g plot the deviations from the grand mean e. This approximates the PCA algorithm, which reorganizes the variance of the individual feature vectors. As our simplified illustration has only two cases, both patterns merge into a single “pseudo-component” that merely changes sign.

The behavioral data set contained 35 cases whose SR matrices mixed the systematic pattern with the toggling pattern (and other idiosyncratic patterns) in different proportions. The SR projections quantify the degree to which these two strategies are expressed in the scanpaths of each individual participant. The systematic projection was positively correlated with the Raven scores, whereas the toggling projection was negatively correlated.

Prior studies have attempted to characterize the constructive-matching and response-elimination strategies with more traditional dwell-time variables. The previous high-water mark was set by Vigneau et al. (2006), who reported $R^2 = .51$ (corrected down
to .48) for predicting Raven scores with a linear combination of the matrix time dis-

tribution index (defined in the Method), the number of toggles on easy items, and the

latency on easy items. When applied to our data, however, these variables achieved

a much lower uncorrected $R^2 = .16$ (Table 2.1). The most that can be achieved with

linear regression on any 3 dwell-time predictors on our data is $R^2 = .21$ (Table 2.1,

bottom row).

Apparently, as Vigneau et al. (2006) acknowledge, these methods of quantifying

eye-movement data are noisy and thus susceptible to over-fitting. This begs the

question of how well the scanpath SR would perform on new data. We conducted

leave-one-out cross-validation to test the generalization performance of our method.

We partitioned the data into a training set of 34 participants and a test set of 1

participant. We ran our two-tier algorithm on the training set. The parameters $\alpha$

and $\gamma$ optimized on the training set were then used to calculate the SR matrix for the

fixation sequences in the test set. Finally, we calculated the model’s prediction of the

test Raven score by multiplying the test SR matrix by the weight matrix estimated

from the training set. We repeated this process 35 times, testing on the data from

each participant in turn. This produced 35 predicted scores, each one based on a

model that had no access to the data that was subsequently used to test it. The

squared correlation between these cross-validated predictions and the observed scores

was $R^2_{cv} = .41$. This is a much better estimate of generalization performance than

the goodness-of-fit $R^2$ on the training set (Haykin, 2009). The latter is inflated

because it reflects not only the genuine regularities in the population, which will

generalize to new cases, but also the idiosyncrasies of the training sample, which

will not. This explains the drop from $R^2 = .56$ to $R^2_{cv} = .41$. Note that this still

is very respectable cross-validated performance, which sets a new benchmark for

Raven score prediction. For comparison, the corresponding values for the best model
Variables used by Vigneau et al. (2006):

- Proportional time on matrix (PTM) \( R^2 = .17 \), \( R^2_{cv} = .09 \)
- Latency to first toggle (FT) \( R^2 = .02 \), \( R^2_{cv} = .01 \)
- Latency on easy items (LEz) \( R^2 = .11 \), \( R^2_{cv} = .04 \)
- Number of toggles on easy items (NT) \( R^2 = .01 \), \( R^2_{cv} = .00 \)
- Toggle rate on easy items (TR) \( R^2 = .12 \), \( R^2_{cv} = .04 \)
- Matrix time distribution index (MTDI) \( R^2 = .02 \), \( R^2_{cv} = .01 \)
- Vigneau et al. model (MTDI+NT+LEz) \( R^2 = .16 \), \( R^2_{cv} = .03 \)
- Best traditional model (PTM+TR+LEz) \( R^2 = .21 \), \( R^2_{cv} = .09 \)

Transition probability matrices (with PCA):

- First-order transitions, 2 components \( R^2 = .29 \), \( R^2_{cv} = .01 \)
- First-order transitions, 4 components \( R^2 = .51 \), \( R^2_{cv} = .07 \)
- Second-order transitions, 2 components \( R^2 = .42 \), \( R^2_{cv} = .19 \)
- Second-order transitions, 4 components \( R^2 = .57 \), \( R^2_{cv} = .26 \)

Table 2.1 Goodness-of-fit \( R^2 \) and leave-one-out cross-validated \( R^2_{cv} \) for predicting Raven scores from eye-movement data. The top line reports the performance of the novel method based on successor representations and principal component analysis (PCA). It is compared to some prominent dwell-time variables from the literature (Vigneau et al., 2006) and to first- and second-order transition probability matrices.

Based on dwell-time variables were \( R^2 = .21 \) and \( R^2_{cv} = .09 \) (Table 2.1). This suggests that the SR algorithm can extract reliable regularities from the data much better than traditional dwell-time methods. The SR advantage comes from the sequential
information in scanpaths and from the data-smoothing properties of the temporal-difference learning algorithm.

The success of the scanpath SR in cross-validated prediction is also a direct result of the stability of the principal components across folds. The same two components—systematicity and toggle—were chosen on all 35 cross-validation folds and were qualitatively indistinguishable from iteration to iteration. Although it is difficult to quantify the component overlap across folds because the two components sometimes switched places, the weight matrices derived from them can be combined linearly. The average weight matrix is shown in Figure 2.4a and is virtually identical to the weight matrix from the global model trained on all data (Figure 2.2c). This suggests that the components were not driven by outliers and reflect genuine dimensions of individual differences in scanpath patterns across the majority of observers. The optimized SR parameter values were also quite stable across the 35 folds: mean $\alpha = .236$ (SD=.02), mean $\gamma = .259$ (SD=.05). The stability of the temporal discount factor $\gamma$ suggests that the scanpath patterns have regularities with a characteristic time scale.

Finally, we compared the new scanpath SR method to first- and second-order transition probability matrices (Table 2.1). We began by calculating the first-order transition matrix for each sequence. Averaging across 28 trials produced one $10 \times 10$ matrix per participant. After reshaping, the first-order data set occupied a matrix of the same size ($35 \times 100$) as the SR data set and was analyzed and cross-validated in the same way. The first 20 principal components retained 89% of the variance in the first-order data. Hierarchical linear regression with 2 components yielded $R^2 = .29$ on the full training set but did not cross-validate ($R^2_{cv} = .01$). Adding variables to the regression model improved the fit only marginally (e.g., $R^2_{cv} = .07$ with 4 components). This suggests that first-order transition matrices are too myopic to support robust prediction of Raven scores. It also demonstrates that the excellent performance of
Figure 2.4 *Leave-one-out cross validation results.* The average weight matrix (a) across 35 leave-one-out fits is virtually identical to the weight matrix produced by the fit to all data at once (Figure 2.2c). Each Raven score was predicted by a separate model that had no access to the data for the respective individual. The squared correlation between the cross-validated predictions and the observed scores was $R^2_{cv} = .41$.

the SR method cannot be attributed to the PCA-based dimensionality reduction algorithm.

Second-order transition probabilities are conditionalized on the two preceding fixations in the sequence. This calls for the estimation of a $10 \times 10 \times 10$ matrix per trial. Given that the median (clipped) sequence length was only 88, the second-order estimates were extremely variable even after averaging across the 28 trials. Still, it was interesting to check whether the PCA algorithm could identify individual differences among the participants. After reshaping, the second-order data set occupied a matrix of size $35 \times 1000$ and the first 20 principal components retained 74% of the variance. Hierarchical linear regression with the second-order projections yielded good fits to the Raven scores (Table 2.1). The best generalizability ($R^2_{cv} = .26$) was achieved with 4 predictor variables. While quite respectable and much better than the $R^2_{cv}$ achievable with traditional measures, this falls far short of the SR-based
prediction. Moreover, unlike the SR-based components (Fig. 2.2), the second-order components were extremely hard to visualize and interpret.

The transition-based results suggest two conclusions. First, a single-step event horizon cannot capture the statistical regularities in our data. A temporally extended analysis seems necessary. This explains why the second-order model performed better than the first-order one. The SR-based model performed the best, due in large part to its open-ended event horizon whose effective size was controlled adaptively by the $\gamma$ parameter. The second conclusion is that the probability estimates need to be smoothed. There are not enough data to populate the matrices by simple counting, particularly in the second-order case. This scarcity of data (rather than computational constraints) appears to be the limiting factor in scanpath analysis in general. The SR learning algorithm (Eq. 2.2.1) updates a whole column of the matrix after each transition, thereby smoothing the estimates. Stated differently, each cell in the SR matrix aggregates a whole class of observations. For example, cell (1,1) would be updated after observing any of the following subsequences: 121, 131, ..., 1R1; 1231, 1241, ... This reuses the data and reduces the variance of the estimates. This smoothing effect contributed to the stability of the SR components during leave-one-out cross-validation. By contrast, the first-order probability estimates were apparently noisier, and the PCA solution was unstable even though it involved matrices of the same shape estimated from the same data.

2.3 General Discussion

Our novel method of eye-movement analysis, the scanpath successor representation (SR), produced new results in terms of both successful score prediction and insight
into individual differences in problem-solving strategies on Raven’s Advanced Progressive Matrices. With this method, we were able to extract the underlying structure from complex patterns of sequential eye movements during geometric problem solving. These regularities allowed us to predict APM scores with unprecedented accuracy. More importantly, the principal-component analysis of the successor representations produced components that were readily interpretable and consistent with earlier strategy findings.

The two components of the scanpath SRs that correlated strongly with the scores mapped clearly onto the two main processing strategies for multiple-choice matrix-completion problems. The anti-toggle component (Figure 2.2b) replicated earlier reports of negative correlations between toggling and Raven scores (Carpenter, Just, & Shell, 1990; Bethell-Fox et al., 1984; Vigneau et al., 2006). This qualitative agreement with established results validates the new SR-based method. Quantitatively, however, it goes a step further because it could predict a larger proportion of the variance compared to traditional measures such as the number of toggles or toggle rate. This suggests that the SR-based analysis provides a more sensitive measure of toggling and thus can better identify individuals who follow the response-elimination strategy. This article did not address the question of whether response elimination is adopted at the beginning of a problem or only as a fall-back strategy on difficult items. This question can be answered by analyzing the SR matrices for individual trials and/or contrasting the early and late portions of the fixation sequences within trials.

The systematicity component (Figure 2.2a) is a novel finding and arguably provides the most detailed picture of Raven performance and strategic processing to date. This component demonstrates the importance of processing the problem matrix row by row. 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 skipping over cells (1→3, 4→6). This suggests that row scanning (particularly adjacent cell scanning within rows) is more likely to generate relational insight, which conforms to previous findings that perceptual-motor patterns can increase the likelihood of rule insight (Grant & Spivey, 2003; Thomas & Lleras, 2007). This lends new support to the theory that successful Raven solvers use a constructive matching strategy and explicates some important aspects of this strategy.

We chose Raven APM as the testbed for the novel scanpath SR method because decades of painstaking research have identified the two strategies most relevant for this domain (e.g., Bethell-Fox et al., 1984; Carpenter et al., 1990; Snow, 1980; Vigneau et al., 2006). Thus, we knew what to expect and could validate the method against these established findings. Still, the method revealed previously unknown details about the constructive matching strategy. More importantly, armed with this powerful tool we could have discovered these two strategies even if we had never read the Raven literature, simply by interpreting the component matrices in Figure 2.2. Note that these matrices were calculated in an entirely automated manner and reflect regularities in the data rather than the prior knowledge of the authors. Thus, the scanpath SR method promises to be a great tool for exploratory data analysis, with the potential for rapid discoveries in other domains.

The power of the scanpath SR stems from the fact that it extends the event horizon of sequential eye movements to extract temporally extended patterns. It will very likely prove useful in any complex task environment that has distinct areas of interest (statically or dynamically defined). This includes other abstract, rule-governed environments such as chess (Charness, Reingold, Pomplun, & Stampe, 2001) or Tower of Hanoi (Patsenko & Altmann, 2010), but also practical applications such as identifying successful and unsuccessful strategies for landing a plane (Anders, 2001;
Ottati, Hickox, & Richter, 1999) or driving a car (Crundall, Underwood, & Chapman, 1998).
CHAPTER 3

Do we really become smarter when our fluid-intelligence test scores improve?

3.1 Introduction

Can intelligence be improved with training? For the most part, the numerous training methods attempted through the years have yielded disappointing results for healthy adults (e.g., Detterman & Sternberg, 1982). Nonetheless, if an effective training method could be designed, it would have immense practical implications. Therefore, when Jaeggi et al. (2008) recently published some encouraging experimental results, they were greeted with remarkable enthusiasm. Cognitive enhancement is now a billion-dollar industry (“Brain sells,” 2013). Millions of customers buy “brain building” games and subscribe to “mental gyms” on-line where they perform various “cognitive workouts” in the hope of raising their IQ (Hurley, 2012). Hundreds of millions of dollars are being invested in educational (e.g., Cogmed, http://www.cogmed.com), military, and commercial programs (e.g., Lumosity, http://www.lumosity.com) on the assumption that intelligence can be improved through training. But can it really? Given the massive societal resources that are at stake and the checkered track record of similar initiatives in the past (e.g., Detterman & Sternberg, 1982; Melby-Lervåg & Hulme, 2013; Owen et al., 2010), this claim must be evaluated very carefully. Here

The text of this chapter [3] is taken from Hayes, Petrov, & Sederberg, in press
we present novel evidence that suggests reasons for skepticism. The evidence is not definitive and the question remains open. It leads directly to three other questions: (i) What is intelligence? (ii) How can we measure intelligence? (iii) How can we measure gains of intelligence? The first two of those have been debated and researched for over a century (see, e.g., Neisser et al., 1996, for an authoritative review). The last question, however, has not received the attention it deserves. One goal of this article is to point out how methodologically challenging it is to measure the change of a latent variable.

With respect to the first two questions, we adopt the popular (though not universally accepted) psychometric approach that both defines and measures fluid intelligence as the latent variable explaining the intercorrelations in performance on tasks such as analogy making, reasoning, and problem solving. This approach is grounded in the fact that individual differences in performance across a wide variety of cognitive tasks are positively correlated (Spearman, 1927). Through factor analysis, the matrix of intercorrelations can be explained in terms of a hierarchical arrangement with a general intelligence factor $G$ at the apex and various more specialized abilities arrayed below it (Carroll, 1993; Jensen, 1998). The second tier in the hierarchy includes the distinction between crystalized ($G_c$) and fluid ($G_f$) intelligence (Cattell, 1963; Carroll, 1993). $G_c$ refers to overlearned skills and static knowledge such as vocabulary, which undoubtedly accumulate with experience. In contrast, $G_f$ refers to the ability to detect patterns and relations, solve problems, and “figure things out” in novel environments. Empirically, fluid intelligence predicts many forms of achievement, especially school achievement (Gottfredson, 1997). There is strong evidence that $G_f$ is highly heritable—between 50% and 75% of the variance of intelligence test scores in healthy adults is linked to genetic variation (Neisser et al., 1996). Although
heritability does not entail immutability (Dickens & Flynn, 2001), most psychometricians conceptualize Gf as a stable trait that is relatively immune to interventions in adulthood (Carroll, 1993; Jensen, 1998).

This is why a recent study by Jaeggi et al. (2008) triggered such excitement and controversy. The study used a pretest-train-posttest design with an untrained control group. A titrated, adaptive dual n-back task was practiced for up to 18 sessions in the experimental group (N = 34) but not in the control group (N = 35). All participants were pre- and post-tested on two parallel short-form versions of a matrix-based Gf test—either Raven’s Advanced Progressive Matrices (Raven et al., 1998) or BOMAT (Hossiep et al., 1999). Whereas the results showed statistically significant score gains in both groups, the average gain in the trained group was significantly higher than that in the control (p < 0.05, η² = 0.07, Jaeggi et al., 2008). The latter finding—a significant control-adjusted gain—was interpreted as an improvement in Gf and fueled the current boom in the cognitive enhancement industry, as well as a big controversy in the scientific literature. Of particular relevance to the controversy is that the original study (Jaeggi et al., 2008) had various methodological shortcomings (Moody, 2009) and subsequent attempts to replicate the putative improvement in Gf have produced mixed results (e.g., Chooi & Thompson, 2012; Harrison et al., 2013; Jaeggi et al., 2011, 2010; Redick et al., 2012; Thompson et al., 2013). This rapidly growing field is characterized by large variations in reported effect sizes (see Melby-Lervåg & Hulme, 2013, for a meta-analysis of 23 studies), polarization of opinion, and contradictory reviews (e.g., Buschkuehl & Jaeggi, 2010; Morrison & Chein, 2011, on the optimistic side; Melby-Lervåg & Hulme, 2013; Shipstead et al., 2012, on the skeptical side).

The neurobiological interpretation of Gf (M. Anderson, 2005; Duncan et al., 2000) emphasizes its linkage to factors such as processing speed (Jensen, 2006; Sheppard &
Vernon, 2008) and working memory capacity (Fry & Hale, 2000; Gray & Thompson, 2004; Halford et al., 2007; Kane & Engle, 2002). The interest in the latter linkage surged after Jaeggi et al.’s (2008) publication because their participants trained on a WM task. The hypothesis that fuels the current enthusiasm is that WM training increases WM capacity (near transfer), which in turn improves Gf (far transfer). There is a strong analogy with athletics, where swimming workouts, for example, increase cardiovascular capacity, which in turn improves the general athletic ability. Thus, Jaeggi et al. (2011) characterize WM as “taking the place of the cardiovascular system.”

This hypothesis is simple and elegant but the methodology for testing it empirically is fraught with difficulties because an objective method for measuring Gf gains is required. The commonly used test-retest method is seriously flawed. The overwhelming majority of studies use test-retest score gains to measure Gf gains. This practice is based on the misleading intuition that if a test such as Raven’s APM is a valid measure of Gf, then a gain in the score on this test is a valid measure of Gf gain. This is not necessarily true because, in addition to Gf, the scores reflect non-Gf factors such as visuospatial ability, motivation, and test-taking strategy. The latter factors—and hence the test scores—can improve while Gf itself remains stable. Indeed, Raven’s APM scores increase significantly on repeated testing without any targeted training (e.g., Bors & Vigneau, 2003; Bors & Forrin, 1995; Denney & Heidrich, 1990). Worse, a large meta-analysis of 64 test-retest studies (te Nijenhuis, van Vianen, & van der Flier, 2007) indicates a strong negative correlation between score gains and the G loadings of test items. To control for such “mere retest” effects, the common practice in the field is to compare the score gains in the treatment group to those in an untreated control group. Cognitive enhancement advocates (e.g., Jaeggi et al., 2008) acknowledge the interpretive problems of unadjusted score gains.
but assume that control-adjusted gains necessarily measure real gains in *Gf*. As we argue below, however, this assumption is incorrect because the adjustment does not guarantee validity either.

These methodological difficulties can be illustrated by analogy with athletics. In a classic study of motor skill learning (Hatze, 1976), an athlete practiced kicking a target as rapidly as possible. His performance improved at first and then plateaued. However, after seeing a film about kicking technique, the athlete immediately improved his time considerably and with additional practice was able to reach a much higher asymptote. For our purposes, this illustrates the relationships between the following three variables. The first is kicking time, which was the only objective measurement. The second variable is general athletic ability, which includes factors such as cardiovascular capacity, agility, muscle strength, and so forth. The third is kicking technique—the optimal way to execute a kick so as to minimize kicking time, all else being equal. Importantly, because the kicking time reflects a mixture of athletic ability and technique, gains in kicking time can occur without any change in athletic ability. Indeed, watching a movie could not have changed the strength or agility of the participant in Hatze’s (1976) experiment. Analogously, gains in test scores can occur without any change in “brainpower” factors such as WM capacity or processing speed.

This brings us to the central topic of transfer across tasks. The most widely used inference pattern in the cognitive enhancement literature is to infer gains in *Gf* on the basis of control-adjusted gains in test scores. This inference pattern logically requires the auxiliary assumption that *only Gf* can transfer across tasks. Few cognitive-enhancement advocates would endorse such a strong claim, and the more cautious authors explicitly disavow it, often near the end of their Discussion sections.
(e.g., Morrison & Chein, 2011, p. 58). But without this assumption, there is no logically necessary link from the observed control-adjusted score gains to the theoretical conclusion of Gf gains. Why not? Because non-Gf-related factors can transfer across tasks too.

The athletic analogy can easily be extended to illustrate this. Suppose that instead of watching a movie, the athlete in Hatze’s (1976) experiment practiced a seemingly unrelated task such as high jump. The problem is that tasks that seem unrelated on the surface can still share critical technical components. For example, the approach of the high jump may actually be as important as the take off. It requires the right amount of speed and the correct number of strides—factors that affect kicking too. So, if an athlete practices high jump for many hours and then can kick a ball faster than before, is this because the jumping practice improved the explosive power of their leg muscles? Or is it because it provided an opportunity to learn to control the approach better? In other words, was there transfer of athletic ability, of technical components, or both? These possibilities cannot be differentiated on the basis of measured gains in kicking speed alone. Analogously, a control-adjusted gain on an intelligence test may stem from genuine Gf transfer from the training task, from transfer of some non-Gf-related component(s), or from a combination thereof.

Despite these interpretive problems, the research community continues to explore various combinations of treatment tasks, control tasks, and tests (see Morrison & Chein, 2011; Melby-Lervåg & Hulme, 2013, for recent reviews), and in many studies the only dependent variable is the (adjusted) gain in test scores from pretest to posttest. This approach treats the test as a black box and yields very few data points per participant, which exacerbates the practical difficulties inherent in multi-session between-subject designs. Progress has been slow and the results have been inconsistent and open to conflicting interpretations as referenced above. In the final
analysis, the problems persist because no conclusive inferences can be drawn on the basis of test-retest comparisons alone. A richer data source is needed.

There are two complementary ways to marshal more data to test whether WM training improves $G_f$. The first is to assess $G_f$ not with a single test but with a broad battery of multiple tests. The second approach is to use tools from cognitive psychology to open the black box and investigate the actual processes that determine the test scores and the gains thereof. In this article we follow the second approach. The topic of multiple tests is introduced only briefly here and will be discussed in more detail later. This literature is in active development and the results are still tentative. Two emerging patterns are particularly relevant to the present analysis. First, when a battery of multiple $G_f$ tests was administered before and after WM training, strong inter-test correlations were found as expected, and yet only some tests showed a significant control-adjusted transfer effect (Colom et al., 2013; Harrison et al., 2013; Jaeggi et al., 2014; Stephenson & Halpern, 2013). This selectivity of transfer highlights that test scores and gains can index distinct aspects of the variability across individuals. The high inter-test correlation presumably reflects the shared $G_f$ loading of scores, whereas the dissociable gains suggest plasticity in one or more non-$G_f$-related factors. This dissociation reinforces the methodological caveats discussed above. The second pattern that emerges from the recent literature is that the tests that did show significant control-adjusted transfer were tests with a prominent visuospatial component\(^1\) (Colom et al., 2013; Jaeggi et al., 2014). This raises the possibility that the experimental intervention in these and earlier studies (e.g., Jaeggi et al., 2008) may have improved the visuospatial ability rather than the

---

\(^1\)By contrast, the scores on verbal tests of $G_f$ did improve from pre- to posttest in both studies, but the gains in the experimental and control groups were statistically indistinguishable.
fluid intelligence of the participants, via the visuospatial demands of the dual n-back task intended for WM training (Moody, 2009; Stephenson & Halpern, 2013).

In this article, we focus on Raven’s Advanced Progressive Matrices (APM, Raven et al., 1998) as the paradigmatic example of the class of matrix-based visual analogy tests that are commonly used in cognitive enhancement research (Buschkuehl & Jaeggi, 2010). A Raven problem consists of a matrix and 8 response alternatives. There are multiple distinct relations among the entries in a given row or column (Figure 3.1, left). To answer the problem correctly, the participant must identify the relations and select the response that matches the pattern. This requires relational reasoning, pattern matching, working memory, executive control, and other abilities central to fluid intelligence. However, Raven scores also depend on test-specific factors, including a prominent visuospatial component. These factors are unrelated to \( Gf \) and are potential confounds in cognitive enhancement research. Thus, it is important to understand them, find ways to measure them, evaluate their potential to contaminate the assessment of \( Gf \) gains, and correct this contamination.

In this article we open the black box of Raven’s APM with the help of detailed eye-tracking data and a novel method for scanpath analysis (Hayes et al., 2011). This rich data source allows us to investigate the information-processing mechanisms associated with the observed gain in test scores. Arguably, this variable—the score gain on a matrix reasoning test—is the most frequently used and potentially misunderstood dependent measure in cognitive enhancement research.

Recently we (Hayes et al., 2011) demonstrated that approximately 40% of the variance of Raven’s APM scores across participants can be predicted on the basis of individual differences in eye-fixation patterns. Critical for this success was a novel data-processing algorithm called \textit{Successor Representation Scanpath Analysis} (SRSA, Hayes et al., 2011) that captures the statistical regularities of \textit{scanpath} sequences of
arbitrary lengths. SRSA uses temporal difference learning (Sutton, 1988) to represent these regularities by a fixed-size matrix called a *successor representation* (SR, Dayan, 1993) that can be aggregated across trials and analyzed with standard multivariate methods such as principal component analysis (PCA, Everitt & Dunn, 2001). Importantly, the SRs are interpretable: Different test-taking strategies give rise to characteristic SR patterns that can be traced in the human data (Figure 3.2). SRSA thus provides unprecedented insight into the role of strategic processing in matrix reasoning tests.

Our goal in this article is to apply this powerful new tool to investigate whether *strategy refinement* can account for the test-retest improvement of Raven scores. The answer is a clear yes. We observed a highly significant practice effect, replicating published results (Bors & Vigneau, 2003; Denney & Heidrich, 1990). Approximately 30% of the variance of score gains across participants could be predicted on the basis of individual differences in the changes in eye-fixation patterns as captured by SRSA. Moreover, the latter changes had a clear interpretation in terms of strategy refinement: Individuals that moved toward a more systematic scanning pattern at posttest also tended to improve their scores. Furthermore, when the strategy-dependent variance was partialed out, the residual score gains were no longer statistically distinguishable from zero. These results indicate that strategy is a critical latent variable and a strong potential confound that must be considered whenever matrix reasoning tests such as Raven's APM are used to measure fluid intelligence gains.

### 3.2 Method

Thirty-five university students with normal or corrected-to-normal vision completed two short-form tests from Raven's Advanced Progressive Matrices, Set II (Raven et al., 1998) on two separate days approximately a week apart. The participants were
paid $6 per hour plus $1 bonus for each correct answer. Half of them completed items 2, 4, 6, 9, 10, 11, 16, 17, 19, 21, 23, 24, 26, and 29 on the first session and 1, 3, 5, 7, 12, 13, 14, 15, 18, 20, 22, 25, 27, and 28 on the second. The other half completed the same subsets in the opposite order. The instructions followed the Raven APM Manual guidelines for individual test administration (Raven et al., 1998). Between the two test sessions, 23 participants completed two additional sessions of paper-and-pencil training on Raven-like problems (Matzen et al., 2010). The remaining 12 participants were no-contact controls.

Each trial began with a brief alert sound. A fixation cross appeared on a 21” CRT monitor in a darkened room (Figure 3.1, right). After the participant fixated for 1 s, the Raven problem appeared and remained onscreen until the participant selected a response using the mouse. Eye-tracking data were collected on both test sessions using a desktop Eyelink 1000 tracker (SR Research, 2006). Saccades and fixations were segmented with Eyelink’s standard algorithm using velocity and acceleration thresholds (SR Research, 2006). Each fixation was assigned to one of 10 distinct areas of interest (AOIs, see Figure 3.1 for details). A single AOI (labeled R) covered the entire response area so that the spatial layout of the answers could not be used to decode the participants’ choices. The few (<1%) fixations outside the 10 designated AOIs were ignored.

3.2.1 Relational Item Scoring

Most APM items contain multiple distinct relations that must be extracted to arrive at the correct answer (Carpenter et al., 1990). However, it is often the case that even when items are answered incorrectly the participant still extracts some of the correct...
Figure 3.1 Example of the Raven’s problem format, relational coding, and trial sequence.

Left: The problem matrix and the 8 response alternatives are shown with solid lines. The height of the rectangular box around the matrix subtended 9 degrees of visual angle. Eye fixations were assigned to 10 areas of interest (AOIs): nine for the matrix cells (top row = 1–3, middle = 4–6, bottom = 7–9) and one for the entire response area. This example item (generated by the authors) requires the extraction of three relations: distribution of three shapes (diamond, triangle, parallelogram), distribution of three line orientations ($0^\circ$, $45^\circ$, $90^\circ$), and quantitative pairwise progression of line numbers ($3 \rightarrow 2 \rightarrow 1$). The vectors above each response were not shown to participants but illustrate the respective relations captured in each possible response. Right: Each trial had three phases: fixation, solution, and response. Eye movements and verbal protocols were collected during the solution phase. Moving the mouse cursor out of the fixation box triggered the response phase, during which the problem matrix was masked and the participant clicked on their chosen answer. The inter-trial interval (ITI) was 200 ms.
relations. On items in which incorrect answers captured some of the correct relations, we used that information to infer which relations were successfully extracted by the participant and were able to increase statistical power by capturing this information. Seven relational rules were identified within the APM items: the five rules introduced by Carpenter et al. (1990) plus two new rules, opacity and unique:

- Constant in a row (CIR): Relation in which an element is the same across rows, but changes down columns.
- Quantitative pairwise progression (PP): Relation in which an element increases or decreases down rows or across columns.
- Figure addition or subtraction (ADD/SUBTRACT): Relation in which an element from one column is added or subtracted from another column to produce a third column element.
- Distribution of three values (D3): Relation in which three values from a categorical attribute are distributed across a row or column.
- Distribution of two values (D2): Relation in which two values from a categorical attribute are distributed through a row, and the third value is null.
- Opacity (OPACITY): Relation indicating which figural elements occlude other figural elements when elements overlap.
- Unique (UNIQUE): Used to demarcate special relations that are specific to an individual APM item.

For every item, each of the eight responses were scored as a vector indicating whether they contained a given relation (1) or did not³ (0). See Figure 3.1 for an example item.

³Four items (11,14,18,27) had responses where partial credit was awarded for relational capture.
coding and Appendix B.1 for the complete relational coding scheme. With this form of relational coding, the participant’s performance for each session was measured as the total number of relations extracted (i.e., the sum of their response vectors) during pre- and posttest, respectively.

3.2.2 Successor Representation Scanpath Analysis

We used SRSA (Hayes et al., 2011) to assess changes in participant strategy by quantifying individual differences in pre- and posttest eye-fixation patterns. SRSA quantifies regularities in sequences of eye-fixations using temporal-difference learning (Sutton, 1988) to construct a matrix called a successor representation (SR, Dayan, 1993). The key idea behind SRSA is that upon observing a transition from one AOI to another, instead of simply updating the transition probability from the first to the second AOI, we associate the first AOI with the second AOI and all expected subsequent AOIs based on prior visits to the second AOI. In this way the SRSA algorithm learns to predict future scanpaths based on past scanpaths. After traversing the entire fixation sequence for a trial, the resulting SR can be conceptualized as having extracted the statistical regularities in temporally extended scanpaths. Specifically, an SR matrix contains, for each AOI, the temporally discounted number of expected future fixations to all AOIs (Dayan, 1993). Given their uniform size and that they are based on the same set of AOIs, the SR matrices from different observers and/or trials can be analyzed using standard statistical methods to identify significant pattern regularities for various comparisons of interest. Since we were interested in examining the change in strategy between pre- and posttest, our present approach was to use the differences between the pre- and posttest SRs to predict the difference between pre- and postest Raven performance.

The first step in SRSA is to convert each trial scanpath into a trial SR. Each
trial scanpath was defined as the sequence of fixations across the 10 distinct AOIs (9 cells of the problem matrix and the response area) on a given trial. A successor representation (Dayan, 1993) was calculated for each trial scanpath, resulting in one $10 \times 10$ SR matrix $M$ per trial for each participant. Each trial SR matrix is initialized with zeros and then updated for each transition in the scanpath sequence. Consider a transition from state $i$ to state $j$. The $i$th column of the matrix—the column corresponding to the “sender” AOI—is updated according to Equation 2.2.1 in Chapter 2: where $I$ is the identity matrix, each subscript picks a column in a matrix, $\alpha$ is a learning-rate parameter ($0 < \alpha < 1$), and $\gamma$ is a temporal discount factor ($0 < \gamma < 1$). The learning rate parameter $\alpha$ controls the incremental updating and $\gamma$ controls the amount of temporal discounting. The latter term is the key to extending the event horizon to encompass both immediate and long-range transitions—it includes the discounted future states in the prediction from the current state. For example, suppose a participant scans the top row of a Raven problem systematically from left to right: $1 \rightarrow 2 \rightarrow 3 \rightarrow 1 \rightarrow 2 \ldots$ Then the successors of location 1 will include both location 2 and, weighted by $\gamma$, location 3. After traversing the whole scanpath, the estimated SR matrix approximates the ideal SR matrix, which contains the temporally discounted number of expected future fixations on all AOIs (rows), given the participant just fixated on any individual AOI (column). Note that the entries in the SR matrix are not probabilities, they are (discounted, expected) numbers of visits. When $\gamma$ is set to zero the SR is equivalent to a first-order transition matrix and as $\gamma$ increases the event horizon is extended farther and farther into the future. Note also that the learning parameter $\alpha$ does not reflect a cognitive learning rate, but only the learning rate that optimizes the temporal-difference learning algorithm.

4 Despite wide variability in sequence length, no sequence clipping (Hayes et al., 2011) was used to attempt to regularize the sequence length for the SRSA difference analysis.
The second step in SRSA depends on the question of interest—in our case the contribution of strategy to Raven APM improvement. Since we were interested in examining strategy differences between sessions, the trial SRs were not averaged across both sessions as they were in Hayes et al. (2011). Instead, for each participant a difference SR matrix was computed by averaging across the session 1 trial SRs and session 2 trial SRs separately, and then taking their difference (mean session-2 SR minus mean session-1 SR), resulting in 35 participant difference SRs. Conceptually, each $10 \times 10$ difference SR captured the difference in eye-fixation patterns between pre- and posttest for the corresponding participant. To reduce the dimensionality of this 100 feature space and prevent over-fitting, we performed a principal-component analysis (PCA, Everitt & Dunn, 2001) of the difference SRs.\textsuperscript{5} PCA is a standard machine learning technique for reducing dimensionality by finding the most informative viewpoints (i.e. variance-maximizing orthogonal rotations) of a high-dimensional space. The result is a set of linear orthogonal variables called principal components. Conceptually, the principal components of the SR differences represent dimensions of individual differences in fixation patterns between pre- and posttest. These are expressed mathematically as orthogonal basis vectors in the 100-dimensional difference SR space. Each participant was characterized by 20 projections onto this rotated basis. The difference SR projections were then used as predictor variables in a multiple linear regression analysis to predict relational score gain (i.e. the difference in the number of relations extracted, posttest minus pretest).

The final step in SRSA is to optimize and cross-validate the model fit between the difference SR projections and relational score gain. We implemented a two-tier algorithm to maximize the fit. In the inner loop, it calculated the difference

\textsuperscript{5}Following standard PCA practice, we re-scaled each feature so that it had zero mean and unit variance across the 35 participants.
SRs for given parameters $\alpha$ and $\gamma$ (Equation 2.2.1), then calculated the first 20 principal components and the corresponding projections for each participant, picked the three projections that correlated most strongly with the relational score gain, and constructed a linear regression model with these three predictors. In the outer loop, a Nelder-Mead optimization routine searched for $\alpha$ and $\gamma$ that maximized the multiple regression coefficient of the inner-loop model. To guard against over-fitting, we performed leave-one-out cross-validation to test the generalization performance of the two-tier fitting algorithm. We partitioned the data into a training set of 34 participants and a test set of 1 participant. We ran our two-tier algorithm on the training set. The parameters $\alpha$ and $\gamma$ optimized on the training set were then used to calculate the SRs for the fixation sequences in the test set. Finally, we calculated the model’s prediction of relational score gain by multiplying the test set difference SR matrix by the weight matrix estimated from the training set. We repeated this process 35 times, testing on the data from each participant in turn. This produced 35 predicted relational score gains, each one based on a model that had no access to the data that was subsequently used to test it. For all SRSA analyses a cross-validated ($R^2_{cv}$) fit is reported.

3.2.3 Results

The relational scores varied between 13 and 32 (M=26.8, SD=4.6) at pretest and between 16 and 33 at posttest across the 35 participants (M=29.0, SD=3.7). The relational score gain (posttest minus pretest) was 2.2 relations on average and varied across individuals (SD=3.9, min=−4, max=11).

6Note for the subgroup analyses (N=11) a reduced set of 6 principal components were used.

7The number of correctly solved problems increased by 1.5 on average ($t(34) = 3.48, p < .001$).
statistically significant \( t(34) = 3.30, \ p = .001 \) (one-tailed), \( d = .56 \) and consistent with earlier reports of practice-induced effects (Denney & Heidrich, 1990; Bors & Vigneau, 2003). Our effect size \( (d = .56) \) was in the upper half of the range of effect sizes typically reported in the \( G_f \) enhancement literature (Melby-Lervåg & Hulme, 2013). The larger effect size may reflect the increased statistical power of our relational scoring scheme compared to Raven’s standard scoring. Despite this abundant statistical power, the paper-and-pencil training manipulation had no significant effect relative to the no-contact control \( (F(2,32) = .98; \ \text{paper-and-pencil} \ M=2.6, \ \text{SD}=4.2; \ \text{no-contact control} \ M=1.4, \ \text{SD}=3.4) \). Thus even without training, Raven performance increased significantly. This illustrates that the mere test-retest procedure is sufficient to induce score gains even when short test forms are used.

A multiple linear regression was performed using the difference SR projections from the PCA to predict the relational score gain for each participant. Utilizing the two-tier fitting algorithm detailed earlier, the best fit \( R^2 = .56 \) was achieved with three principal components, learning rate \( \alpha^* = .35 \), and discount parameter \( \gamma^* = .29 \). As was shown in Hayes et al. (2011), eye-movement data are susceptible to overfitting and so it is essential to perform leave-one-out cross validation to test the generalization performance. Using cross-validation we were still able to account for approximately a third of the variance in relational score gains from pre- to posttest: \( R^2_{cv} = .32 \). Panel a in Figure 3.2 shows the average prediction weight matrix across the 35 leave-one-out fits and panel b plots the cross-validated predictions against the observed gains. The average prediction weight matrix reflects the sum of the principal components (scaled by their respective regression coefficients) averaged across the 35 leave-one-out fits.

Just as important as the amount of variance explained by the difference SRs is the clear interpretation offered by the prediction weights themselves. The dominant patterns that were observed in the difference SR principal components are reflected in
Figure 3.2 Weight matrices, relational score gain predictions for the full cross-validated model, and Simulated SR differences. The cross-validated model prediction weight matrix across 35 leave-one-out fits (a) revealed a strong relationship between systematic scanning and relational score gains across sessions. The relational score gain was predicted by a separate model that had no access to the data for the respective individual. Panel b plots the predicted versus observed relational score gain for all 35 participants ($R^2_{cv} = .32$). Panel c and d were generated using simulated scanpath sequences to highlight important structure. Panel c shows an idealized difference SR resulting from simulated sequences with a 90% increase in row-systematicity on session 2. Panel d shows an idealized difference SR resulting from simulated sequences with a 20% boost in answer checking on session 2. The x- and y-axes represent the sender and receiver areas of interest, respectively. R = response area of interest. SR = successor representation of the regularities in scanpath sequences.
the prediction weights. In particular, the diagonal box structure indicates systematic row-wise scanning (cf. Figures 3.2a and 3.2c). This finding suggests that a significant portion of the practice effect was associated with refinements in information processing strategy whereby participants scanned rows of the problem more systematically and were less prone to haphazard scanning at posttest. In addition to the diagonal box structure indicative of a constructive matching strategy, the weight matrix in Figure 3.2a also has “hot spots” in the bottom-left and top-right corners. This pattern indicates an increase in the systematic scanning of cells 1 2 3 (top row) followed by cells 8 and 9 (which need completion), followed by inspection of the response area (cf. Figures 3.2a and 3.2d).

To get a clearer picture of the differences between participants that improved and those that got worse, we ran separate cross-validated models for the 11 participants that improved the most and the 11 participants that performed worse or showed no improvement at posttest. For both subgroups, the difference SRSA was able to predict a significant portion of the variance in relational score (low group $R^2_{cv} = .30$; high group $R^2_{cv} = .44$). The average prediction weights across the 11 leave-one-out fits are shown in Figure 3.3. The low group prediction weights shows more diffuse weights with no clear diagonal structure as well as some off-diagonal values, indicative of more haphazard scanning. This means that participants whose scores stayed the same or worsened at posttest used the same or less optimal scanning strategies on session 2 relative to session 1. The high group prediction weights shows the opposite pattern with an even stronger diagonal box structure than the full model, which clearly shows a strategically driven improvement in relational extraction. The bottom-left and top-right weight pattern is also brought into better focus in this high-improvement group. As discussed above, this pattern can be generated from a sequential systematic scanning of the first row, cells 8 and 9, and then the response area (Figure 3.2d). We
Figure 3.3 Comparison of cross-validated prediction weights for high- and low-improvement groups. Relational score gains were predicted separately for the eleven highest and eleven lowest improvement participants across 11 leave-one-out fits (low $R^2_{cv}=.30$; high $R^2_{cv}=.44$). Each value was predicted by a separate model that had no access to the data for the corresponding individual. Panel (a) shows the average prediction weight matrix for the low-improvement group and panel (b) for the high-improvement group. A comparison of the prediction weight matrices shows markedly more diffuse scanning in the low-improvement group (panel a) and a gain in systematicity in the high-improvement group (panel b).

interpret this sequential pattern as an indication that participants are checking their answer more carefully at posttest prior to selecting it. These results are a further demonstration that strategy refinement between pre- and posttest can account for changes (gains and losses) in Raven’s APM performance.

To determine whether our practice effect remained after removing strategic gains in our participants, we performed a residual analysis to determine whether the significant practice effect we observed would survive in the absence of the strategic improvements that are clearly evident from the SRSA analysis. In both the entire
group \((t(34) = .30)\) and even the high-improvement subgroup \((t(10) = .25)\), the practice effect was no longer statistically significant after the SR covariate was partialled out.

### 3.3 General Discussion

In this article we used eye-tracking data and a novel method for scanpath analysis to investigate the information-processing mechanisms associated with practice effects on matrix-based visual analogy tests. The results showed significant test-retest gains in the Raven scores (Bors & Vigneau, 2003; Denney & Heidrich, 1990). Importantly, over 30% of the variance of score gains across participants could be attributed to refinements in problem-solving strategies as revealed by characteristic changes in eye-fixation patterns. Moreover, when the strategy-related variance was partialled out, the residual score gains were no longer significant, even in the high-improvement subgroup. This indicates that strategy refinement is a powerful determinant of score gains—it controls a major portion of the variance and can change the substantive conclusion of an experiment. Consequently, it must be considered carefully when interpreting score gains on Raven’s APM and similar matrix-based tests.

The central topic in the cognitive enhancement literature is the topic of transfer across tasks. We acknowledge that, given the lack of a transfer group in our experiment, our data do not bear directly on this topic. Nevertheless, the present article contributes to this literature in two ways: empirical and conceptual. The empirical contribution is to examine in detail the information-processing mechanisms underlying the most frequently used dependent measure in the \(Gf\) enhancement\(^8\) field—the score gain on a Raven-like matrix test. Until recently (e.g., Buschkuehl & Jaeggi, 2012)

---

\(^8\)Of course, the broader field of cognitive enhancement employs a broad variety of dependent measures.
the overwhelming majority of positive reports of far transfer of WM training to fluid intelligence relied exclusively on control-adjusted score gains on such tests. In effect, our results provide unprecedentedly detailed information on the likely mechanism for the score gains observed in the control groups of these experiments. Further research is needed to investigate whether the same mechanism can account for the gains in the WM training groups as well. The parsimonious hypothesis is that it does, barring evidence to the contrary.

This hypothesis is also consistent with the longstanding distinction between the acquisition of skills and the improvement of abilities (e.g., J. R. Anderson, 2000). The former supports transfer only between tasks that have procedural and/or declarative knowledge in common, whereas the latter implies gains in general mechanisms and capacities that carry the potential for widespread transfer across diverse tasks. The difficulty of achieving such broad transfer has long frustrated educators. Decades of instructional research have demonstrated that it is hard enough to acquire specific skills but much, much harder to improve general abilities (J. R. Anderson, 2000). Given this general pattern, it seems much more likely that the transfer of WM training to Raven-like tests (Jaeggi et al., 2008) is due to skill acquisition—including strategy refinement—rather than improvement of \( G_f \).

The conceptual contribution of this article is to articulate an assumption that is logically required for inferring \( G_f \) gains on the basis of test score gains—namely, that only \( G_f \) can transfer across ostensibly different tasks such as n-back and Raven’s APM. As we argued in the introduction, this assumption cannot be taken for granted because non-\( G_f \)-related factors can transfer across tasks too.

Consider motivation as a case in point: Participants who have invested time and effort to practice a challenging WM task are likely to be more motivated on the posttest compared to control participants. Higher motivation is expected to raise the
test scores in the experimental group even when the treatment has no effect on fluid intelligence. When a suitably chosen “placebo” practice made the control participants equally motivated, their test scores improved by approximately the same amount in some studies (Redick et al., 2012; Melby-Lervåg & Hulme, 2013). Thus, motivation is an example of a factor that can sometimes transfer across different tasks and yet is clearly distinct from Gf. It should be mentioned that some studies (e.g., Jaeggi et al., 2011, 2014) suggest that motivation by itself cannot account for the totality of the improvement on reasoning tests. This fact, however, does not invalidate our general methodological point: It cannot be assumed that nothing except Gf can transfer from a WM task to a reasoning task. This is a substantive hypothesis that must be articulated explicitly and supported experimentally (Harrison et al., 2013; Shipstead et al., 2012).

Our results identify another factor that must be considered carefully: cognitive strategy. This is consistent with the evidence that strategy plays an important role in many tasks (e.g., Pressley et al., 1990; McCormick, Miller, & Pressley, 1989; Sternberg & Weil, 1980), tests (Bond & Harman, 1994), and specifically in Raven’s APM (Bethell-Fox et al., 1984; Carpenter et al., 1990; Hayes et al., 2011; Vigneau et al., 2006). Hence the correlation between strategy refinement and Raven score gains is not too surprising. Nevertheless, it is notable how strong the correlation is and that it accounts for a significant portion of the improvement in test scores.

Raven’s APM includes a significant visuospatial component in addition to its well established Gf component. Jensen (1998) estimates that 64% of the variance in Raven’s scores are attributable to Gf. Other studies (e.g., Kane et al., 2004; Schweizer, Goldhammer, Rauch, & Moosbrugger, 2007) yield similar estimates. Thus, 30–40% of Raven’s variance is not related to Gf. While some of this residual variance is just random noise, some of it is systematic. In the study of Schweizer et al. (2007), for
instance, there was 11% and 7% variance overlap between Raven’s APM and Horn’s (1983) visualization and mental-rotation scales, respectively. This is not surprising given the visual nature of the test (Figure 3.1). Theoretical (e.g., Carpenter et al., 1990), and computational (e.g., Lovett, Tomai, Forbus, & Usher, 2009) models of Raven’s APM also include a prominent visuospatial component. Analogous considerations apply to BOMAT (Hossiep et al., 1999) and all other matrix reasoning tests used in \( Gf \) enhancement research.

It is important to dispel a tempting interpretive mistake that arises at this point. For concreteness, let us assume that 60% of the variance in Raven’s scores are attributable to \( Gf \), whereas less than 10% are attributable to visuospatial ability. One might argue on the basis of these figures that the main \( Gf \) component dwarfs the visuospatial “contamination.” This is the rationale for the widespread acceptance of Raven’s APM as a unidimensional measure of \( Gf \) (Raven et al., 1998). However, these figures apply to Raven’s scores across individuals, whereas the dependent measure in WM training studies is the difference between two scores for the same individual. If \( Gf \) is a stable latent variable, it will contribute equally to the pre- and posttest scores and this contribution, no matter how large, will cancel out in the subtraction. Therefore, \textit{the variance of the score gains can have a radically different composition than the variance of the scores themselves}. Indeed, a meta-analysis of 64 test-retest studies (te Nijenhuis et al., 2007) found a strong \textit{negative} correlation between score gains and the \( G \) loadings of test items.

This illustrates a general limitation of score gains—they can lead to fallacious conclusions and hence must be interpreted with great caution. Some prominent methodologists have even advised against their use altogether: “Gain scores are rarely useful, no matter how they may be adjusted or refined. . . . Investigators who ask
questions regarding gain scores would ordinarily be better advised to frame their questions in other ways” (Cronbach & Furby, 1970, p. 80).

Given that fluid intelligence is defined as the latent variable explaining the intercorrelations in performance on a wide spectrum of tasks (Cattell, 1963; Carroll, 1993; Jensen, 1998; Martínez et al., 2011; Spearman, 1927), one must employ a comprehensive battery of tests to evaluate whether $G_f$ improves with practice at the latent level—that is, “at a level that represents the components of the variance common to the set of tasks indexing a given ability” (Schmiedek, Lövden, & Lindenberger, 2010, p. 2). This methodological imperative is gradually being acknowledged in the field and there is a growing number of studies that administer multiple tests (Colom et al., 2013; Harrison et al., 2013; Jaeggi et al., 2014, 2011; Schmiedek et al., 2010; Stephenson & Halpern, 2013; von Bastian & Oberauer, 2013). As these studies are too complex to review in detail here, we will restrict our discussion to findings related to the topic of visual strategies.

These recent multi-test data suggest the possibility that the putative gain in fluid intelligence may actually be gain in visuospatial ability. The study of Stephenson and Halpern (2013) was designed to test this possibility. It included multiple training groups practicing purely visual, purely auditory, or dual versions of the n-back task. The results showed significant control-adjusted gains on only two out of four $G_f$ tests and only for participants who had a visuospatial component in training. A limitation of Stephenson and Halpern’s (2013) design was that it tested transfer exclusively in the visual modality. By contrast, Jaeggi et al. (2014) included non-visual tests in the battery of outcome measures. The results showed significant transfer on the visuospatial reasoning tests in the visual training and the auditory training group, but no significant transfer on the verbal reasoning tests in either training group relative to the control group. Again, this is consistent with the hypothesis that transfer
might be restricted to the visuospatial domain. Jaeggi et al. (2014) temper this conclusion with the caveat that the verbal reasoning measures have lower reliability and hence afford less statistical power than the visuospatial measures. A third study (Colom et al., 2013) also administered both visuospatial and verbal reasoning tests, and used item response theory (IRT) to derive indices of Gf and other constructs. No statistically significant transfer was obtained for any construct, although there was a trend for Gf ($p < .06$). This trend was undermined by the lack of significant near transfer to the WM construct (cf. Melby-Lervåg & Hulme, 2013; Shipstead et al., 2012). Moreover, once again the Gf transfer was limited to the visuospatial tests, whereas the verbal reasoning test improved equally in both training and control groups. A study with older adults (Stepankova et al., 2014) also found improvement in visuospatial skills following verbal n-back training. Finally, two studies (Schmiedek et al., 2010; von Bastian & Oberauer, 2013) report statistically significant Gf transfer at the latent level. Upon closer examination, however, these data too are compatible with the visuospatial hypothesis because the gain in the latent reasoning factor seems driven by visuospatial tests in both studies. This is hard to evaluate from von Bastian and Oberauer’s (2013) report because it tabulates the results only in terms of an aggregate reasoning score that lumps the verbal and visuospatial modalities together. There is a purely verbal reasoning test—syllogisms—included in the report and the error bars in von Bastian and Oberauer’s (2013) Figure 5 suggest that it did not transfer significantly. We should also note that all data are reported and analyzed in terms of standardized gain scores, which must be interpreted with caution as discussed above. The statistical analysis of Schmiedek et al. (2010) is more sophisticated. It employs a latent difference score model (McArdle & Nesselroade, 1994) that uses factor-analytic techniques to evaluate gains at the latent level. This study compared younger and older adults. The results showed a small (effect size
but statistically significant transfer effect for the $Gf$ latent factor in the younger experimental group (relative to younger control) and nonsignificant transfer in the older experimental group (relative to older control). At the level of individual tasks in the younger group, the greatest transfer in the reasoning category occurred in the visuospatial modality ($d = .38$), whereas the verbal modality showed a trend ($d = .13$) but did not reach statistical significance ($p = .26$). Interestingly, reasoning was also tested in the numerical modality and it did show significant transfer ($d = .33$) in the younger group (Schmiedek et al., 2010, Table 3). The interpretation of these results is complicated by the fact that both experimental groups practiced a diverse array of 12 tasks spanning all three modalities. Thus, it is possible that the transfer to numerical reasoning is driven by training on a numerical task, the transfer to visuospatial reasoning is driven by training on a visuospatial task, etc. Consequently, even this rich data set does not allow definitive conclusions with respect to the aforementioned distinction between the acquisition of skills and the improvement of abilities. In summary, the issues are complex and the results are not easy to interpret. Still, the available multi-test data seem consistent with the hypothesis that the observed $Gf$ gains may be visuospatial gains in disguise.

In conclusion, let us recapitulate the diverse strands of evidence considered in this article. Fluid intelligence ($Gf$) is defined as a latent variable that cannot be measured directly but must be inferred from the intercorrelations in a diverse battery of tests. There is strong evidence that $Gf$ is highly heritable. The prevailing opinion among psychometricians, based on decades of research and disappointments with past efforts at improvement, is that $Gf$ is a relatively stable trait. The recent wave of enthusiasm in $Gf$ enhancement was triggered by reports of score gains on matrix reasoning tests. The interpretation of these results is questionable because no single test score is identical with $Gf$ and because score gains can be dominated by factors
that play marginal roles in the scores themselves. The data reported here show score gains on Raven’s APM that are commensurate with the effect sizes typical of cognitive enhancement studies. Importantly, these gains can be accounted for in terms of refinements in problem-solving strategies as revealed by characteristic changes in eye-fixation patterns. Our data do not address whether the same mechanism can account for the entire transfer of WM training to Raven-like tests. However, the newest studies that assessed $Gf$ via a diverse battery of tests raised the possibility that the transfer may be restricted to the visual modality. This indirectly supports the hypothesis that at least some of this transfer may be driven by refinements in visual scanning strategies. This hypothesis is also consistent with established theories of skill acquisition that explain transfer in mechanistic terms. By contrast, the alternative hypothesis is usually formulated by means of vague analogies with athletics. We are not aware of a mechanistic proposal of how n-back training improves WM capacity. The $Gf$ improvement hypothesis is advanced on the basis of data showing higher score gains on Raven-like tests following WM training compared to control. This inference logically depends on the assumption that $Gf$ gain is the only possible explanation for such control-adjusted transfer. This assumption cannot be taken for granted because non-$Gf$-related factors can transfer across tasks too. Notably, procedural knowledge can transfer in subtle ways even between tasks that seem unrelated on the surface, and especially between overlapping tasks such as visual n-back and Raven’s APM.

On the basis of this converging evidence, we conclude that it is entirely possible, indeed likely, that the reported transfer of WM training to Raven-like tests is due at least in part to refinements in visual scanning strategies. More broadly, the control-adjusted score gains probably include a contribution from procedural knowledge tacitly acquired and fine-tuned during the WM training and later utilized at posttest.

If strategic procedural knowledge transfers across tasks, does WM training induce
Gf gains that cannot be explained in terms of strategic transfer? The remainder of this article outlines some methodological recommendations on how to investigate this question experimentally in the future.

The most informative experimental designs are characterized by two features: focused training interventions in several distinct groups, and pre- and post testing with a comprehensive suite of outcome measures. The study of Jaeggi et al. (2014) illustrates a well designed set of training interventions: one group practiced exclusively the auditory n-back task, a second group practiced the dual (audio and visual) n-back task, and there was also an active control group. As for the outcome measures, it is necessary to assess three types of outcomes for each participant before and after training. First, Gf must be assessed with a battery of tests as discussed above. It is important to include both visual and non-visual reasoning tests in this battery. The tacit assumption that Raven’s APM (or any other test, for that matter) equals Gf is too simplistic. Second, a battery of visual and non-visual WM measures is needed to assess near transfer (Shipstead et al., 2012). Third, the visual scanning strategies must also be assessed, and the tools developed here provide the means to do so. Our data demonstrated that strategy refinement can control a substantial portion of the variance and that, therefore, strategies must be monitored and taken into account in the analysis. We recommend to administer all visual tests with an eye tracker and to process the scanpath data with the SRSA algorithm (Hayes et al., 2011). The resulting successor representations (or, more parsimoniously, the first few principal components thereof) should be included to the suite of outcome measures and used as covariates in the main statistical analysis.

The statistical analysis must estimate latent variables and test whether Gf improves at the latent level (McArdle & Nesselroade, 1994; Schmiedek et al., 2010). We share Cronbach and Furby’s (1970) reservations about score gains as measures
of change, particularly with respect to a variable that is defined at the latent level. Fortunately, quantitative psychologists have developed sophisticated methods for analyzing learning and change at the latent level. A test of the training effect on $Gf$ can be realized by using a bifactor model (Yung, Thissen, & McLeod, 1999) with $Gf$ as the general dimension. The model must guarantee that the nature of the latent variable does not change from pretest to posttest and that the training effect is an effect on this general dimension. One method that guarantees this is the Multiple-indicator multiple-cause (MIMIC) model (Goldberger, 1972) with pretest-versus-posttest as an external covariate of the general dimension that is shared by pretest and posttest. The same modeling framework also makes it possible to estimate effects on more specific latent variables and to isolate a strategy-specific effect from a genuine effect on $Gf$. The Latent difference score model (McArdle & Nesselroade, 1994) is based on similar principles and has similar virtues. It has already been applied successfully to cognitive enhancement data (Schmiedek et al., 2010). A second approach to guarantee comparability between pretest and posttest is to analyze the data at the level of individual test items instead of aggregate scores. Item response theory (De Boeck & Wilson, 2004) can then be used to impose constraints on the item parameters at pretest and posttest. This approach is developed in Embretson’s (1991) model of learning and change.

Empirical research along these lines has the potential to identify which aspects of intelligent performance improve after what kind of practice via what mechanisms. We are aware of the logistical difficulties in collecting so much data per participant, including eye tracking, and latent-level modeling. However, no simpler methodology can overcome the interpretative difficulties inherent in demonstrating change in a latent variable in the presence of intercorrelated confounds, and pinpointing the causes

---

9We thank Paul De Boeck for his expert advice on these methods.
for this change. Given the massive societal resources at stake and the enormous potential benefit, this research burden is clearly warranted.

Finally, we come full circle to our opening question: Can intelligence be improved with training? The issues are complex and much of the current disagreement stems from incompatible interpretations of the vague and ambiguous term “fluid intelligence.” One important piece of this large puzzle is the ability to flexibly deploy a judicious variety of cognitive strategies and to adaptively learn their utilities for various tasks. If this ability is taken to be part and parcel of $G_f$ then the answer to the opening question may well be yes. If, however, $G_f$ is interpreted in narrow neurobiological terms (e.g., Duncan et al., 2000; Gray & Thompson, 2004) then the answer remains elusive. So far we have seen no conclusive evidence that the brain can be trained like a muscle.
CHAPTER 4
Quantifying the role of strategy and attentional scope during relational reasoning

4.1 Introduction

General fluid intelligence (Gf) refers to our novel problem-solving ability and represents a major dimension of individual difference, predicting performance across a wide range of important areas including educational achievement and job performance (Gottfredson, 1997). Yet despite the broad importance of Gf, significant questions still remain about what contributes to individual differences in fluid intelligence. There is some evidence indicating that individual differences in executive function (EF) and working memory capacity (WMC) are correlated with Gf and each other (Kane & Engle, 2002; Unsworth & Engle, 2005; Ackerman, Beier, & Boyle, 2005; Engle, Tuholski, et al., 1999). However, there is still considerable debate on exactly how to define and measure executive function and working memory capacity (See Blair, 2006) and it is fair to say that the source of individual differences in fluid intelligence remains elusive (Garlick & Sejnowski, 2006; Neisser et al., 1996). The interconnected relationship between EF and WMC as well as the wide array of different tasks used to measure each construct has made it difficult to understand the precise contribution of EF and WMC to fluid intelligence. Recent findings have suggested individual differences in attentional control (Engle & Kane, 2004; Heitz & Engle,
and attentional scope (Cowan et al., 2005; Cowan, Fristoe, Elliott, Brunner, & Saults, 2006) may be important to understanding the complex interplay between EF, WMC, and their role in Gf. Here we add new evidence to support an important role for attentional control and attentional scope by actively manipulating the order and amount of problem information during relational reasoning on a benchmark test of fluid intelligence Raven’s Advanced Progressive Matrices (APM, Raven et al., 1998).

The Raven APM is a fluid intelligence test with excellent psychometric properties (Brouwers et al., 2009) that has, for 70 years, been a reliable and trusted instrument for measuring fluid intelligence across clinical (e.g., Soulieres et al., 2009), developmental (e.g., Eslinger et al., 2009), and cognitive (e.g., Gray et al., 2003) psychology. Each Raven APM item consists of a 3x3 problem matrix in which the bottom right cell is missing and participants must select the best answer to complete the problem matrix from 8 possible response alternatives (Figure 4.1, left). As the APM test progresses the relations contained in each new problem matrix increase and become more complex. Extracting relational patterns from this type of novel relational environment requires a broad group of processes implicated in fluid intelligence including working memory, attentional control, strategy, goal management, and relational binding (Carpenter et al., 1990; Bethell-Fox et al., 1984; Oberauer, Sub, Wilhelm, & Sander, 2007; Hayes et al., 2011). There has been considerable debate about the contributions of each of these processes to fluid performance as well as debate on the extent to which they are separable constructs (See Blair, 2006 for a review). The debate has been complicated by intercorrelations between processes, differences in the tasks used to measure each construct, and the shared neural substrate of the component processes, the prefrontal cortex (Gray et al., 2003; Prabhakaran, Smith, Desmond, Glover, & Gabrieli, 1997; Duncan et al., 2000; Duncan, Burgess, & Emslie, 1995; Roca et al., 2010). Each of these factors has contributed to the difficulty in understanding
individual differences in fluid intelligence. However, despite these difficulties, there have been some consistent findings on the importance of working memory capacity and attention.

Working memory capacity is strongly correlated with fluid intelligence (Heitz et al., 2006). A number of latent variable studies have estimated the common variance between WMC tasks and Gf tests like Raven, estimate shared variance between 36% and 64% (Kane, Hambrick, & Conway, 2005; Ackerman et al., 2005; Conway, Kane, & Engle, 2003; Heitz, Unsworth, & Engle, 2004). The range in the estimates of shared variance is driven in part by the type of WMC task used, with lower estimates typically obtained in strict short-term memory (STM) tasks (e.g. digit and backward/forward span tasks) and higher estimates in complex WM span tasks (e.g. operation span and reading span) which require participants to maintain information in memory while also resisting interference from irrelevant information from secondary tasks (Kane & Engle, 2002; Garlick & Sejnowski, 2006; Daneman & Carpenter, 1980). The correlation between fluid intelligence and working memory is not surprising as Gf tasks like Raven necessitate a way to store and maintain information during ongoing processing, but the reason for the correlation is widely debated. Attentional control (Engle & Kane, 2004; Unsworth & Engle, 2007) and attentional scope (Cowan et al., 2005, 2006) have each been suggested as possible mechanisms for the shared variance between Gf and WMC.

Attentional control or the ability to focus attention on goal-relevant information while blocking interfering information is thought to contribute to working memory capacity and has led some to suggest it as the ultimate source of the shared variance between complex working memory span tasks and fluid ability measures (Engle, Tuholski, et al., 1999; Kane & Engle, 2002). Attentional control is typically measured by tasks which require strong inhibition of competing information (e.g. antisaccade,
flanker, and Stroop tasks) and plays a critical role in theories of information processing and working memory (Atkinson & Shiffrin, 1968; Baddeley, 1986). Two pieces of evidence from latent variable analyses reflect the potential importance of attentional control to understanding individual differences in fluid intelligence: (i) WMC and attentional control are correlated yet dissociable (Heitz et al., 2006; Miyake et al., 2000; Miyake, Friedman, Rettinger, Shah, & Hegarty, 2001) and (ii) when WMC is separated into storage and attentional control factors, attentional control is most closely associated with Gf (Engle, Tuholski, et al., 1999; Conway et al., 2002; Engle & Kane, 2004). Therefore, understanding individual differences in fluid intelligence may first require a better understanding of the role of attentional control during relational reasoning.

Strategy also plays an important role in accounting for individual differences in fluid intelligence (Snow, 1980; Bethell-Fox et al., 1984; Carpenter et al., 1990; Hayes et al., 2011). Early findings using eye movements during geometrical analogy performance revealed two broad strategies: constructive matching and response elimination (Bethell-Fox et al., 1984; Carpenter et al., 1990; Vigneau et al., 2006). In constructive matching, the participant tries to formulate the missing element based exclusively on matrix information, and then looks for that element in the response area. In response elimination, each alternative is inspected in turn and evaluated whether it fits into the empty matrix slot. The former strategy tends to occur in high-scoring individuals and/or easier problems, the latter in low-scoring individuals and/or difficult problems. More recent findings using temporal difference learning to analyze eye movement scanpaths revealed individual differences in scanning patterns that explained 40% of the variance in Raven APM scores (Hayes et al., 2011). Specifically, the findings showed that participants that attend to Raven problem information systematically across rows and do not frequently switch between problem and response information tend
to have higher scores. In some sense these broad processing strategies can be thought of as providing high-level representations of how attentional control is deployed over time during relational reasoning (Hayes et al., 2011; Carpenter et al., 1990). These results and others studying the role of eye movements and attention during insight problem solving have suggested that the way one scans relational information may affect the probability of relational insight (Grant & Spivey, 2003; Thomas & Lleras, 2007). Taken in sum, these findings indicate the study of the role of information processing strategy has the potential to offer new insight into individual differences in fluid ability and the contribution of attentional control.

Attentional scope (i.e. the capacity of the focus of attention) has been suggested as a distinct construct that may uniquely contribute to individual differences in fluid intelligence and help decipher the relationships between EF, WM, and Gf (Cowan et al., 2005; Cowan, 2001). Whereas attentional control and strategy reflect how attention is controlled over time to process information, attentional scope reflects the total capacity of the focus of attention at any given time (Broadbent, 1975; Cowan, 2001; Cowan, Chen, & Rouder, 2004). Cowan et al., 2006 recently examined how attentional scope and control related to verbal and nonverbal aptitude measures in children and adults. In their study 52 children (mean age 10.8) and 52 adults (mean age 20.5) completed two attentional scope tasks (auditory digit span, visual letter span), two attentional control/WMC tasks (dual task, visual array task), and two intelligence subtests from the Standord-Binet Intelligence Scale (vocabulary subtest and pattern-analysis subtest) (Cowan et al., 2006). Their results indicated i) attentional scope (not attentional control) predicted intelligence in children and (ii) individual differences in attentional scope and attentional control together predicted 37% of the variance in intelligence in adults. Attention scope uniquely explained 15% of the variance, attention control uniquely explained 10% of the variance, and
approximately one third of the explained variance in intelligence was shared (12%). These findings reinforce the idea that individual differences in attentional scope may be important to understanding individual differences in fluid intelligence.

Attention scope is related to theories of visual attention that have suggested distinct modes of attentional focus (Heitz & Engle, 2007; Jonides, 1983; Eriksen & St. James, 1986). The two-process model proposed by Jonides (1983) argues that attention can be applied in two distinct modes: a diffuse mode in which attentional resources are equally distributed across the visual field or a focused mode in which attentional resources are concentrated at one area of interest. The two-process model was inspired by findings that under some conditions attention can be distributed evenly over a display (Eriksen & Spencer, 1979; Shiffrin & Gardner, 1972) and in others narrowly focused (Hoffman, 1978, 1979). Similar models such as the zoom-lens model argue that the spatial extent of visual attention is dynamic and changes in size along a continuum in relation to task demands (Eriksen & St. James, 1986). A key idea in both these models of visual attention is as visual attention expands its spatial extent, processing slows due to limited attentional resources (Jonides, 1983; Eriksen & St. James, 1986). Therefore, it is important to consider that individual differences in fluid intelligence may not only rely on the capacity of attention, but also may reflect differences in how the attentional “spotlight” is dynamically updated during relational reasoning.

Much of the work to date on the relationship between WMC, EF, and Gf has relied on latent variable analyses to map the relationships between working memory tasks, executive function tasks, and fluid ability tests (e.g., Engle & Kane, 2004; Engle, Tuholski, et al., 1999; Unsworth & Engle, 2005). While this approach has advanced our understanding of the shared and distinct variance between these three latent variables, it is unclear exactly how these variables actively manifest themselves during
relational reasoning and contribute to individual differences in fluid intelligence. A complementary approach is to study how individuals actually perform relational tasks and predict individual differences in fluid intelligence using metrics such as eye movements or verbal protocols that correspond with variables of interest such as strategy and attentional control during relational reasoning (e.g., Bethell-Fox et al., 1984; Carpenter et al., 1990; Hayes et al., 2011). A third approach is to actively target specific aspects of relational reasoning by comparing relational performance under normal testing conditions to conditions that actively manipulate how relational information is presented to observers (e.g., Grant & Spivey, 2003; Thomas & Lleras, 2007).

The two experiments presented here use the third approach to studying relational reasoning on the Raven APM fluid intelligence test and seek to: (1) experimentally establish and quantify the relationship between strategic processing and relational insight and (2) test whether constraining attentional scope affects relational performance. Experiment 1 extended previous findings on the link between strategic information processing and the probability of relational insight by actively manipulating the order of attended Raven APM problem information. A within-subjects design compared relational performance under normal testing conditions to performance when problem information was presented systematically by row, haphazardly, or was freely controlled by participants. The results revealed a clear causal relationship between strategic attentional control and relational insight, with systematic scanning resulting in increased likelihood of relational insight and haphazard scanning producing a deficit in relational insight. Experiment 2 used a gaze-contingent manipulation that reduced the field of view to one cell to study the effects of reducing the amount of information available by actively constraining attentional scope. The large reduction in attentional scope, surprisingly, did not produce significant deficits in APM score or solution time across participants.
Figure 4.1 *Example of the Raven problem format and active manipulation paradigms* Left: The problem matrix and the 8 response alternatives are shown with solid lines. The height of the rectangular box around the matrix subtended 9 degrees of visual angle. This problem was generated by the authors to protect the security of the standardized test. Center: Example of the session 2 viewing conditions in Experiment 1. The problem and response areas were outlined to allow easy orientation within each item. Only one cell of the problem could be viewed at any given time and the order of problem information was either row-wise systematic, haphazard, or free. The blue lines indicate possible systematic transitions given the current cell while the red arrows indicate possible haphazard transitions. In the control condition the participant could freely transition to any cell using the mouse. Right: Example of the session 2 viewing conditions in Experiment 2. Participants viewed each item through a gaze-contingent viewing window (diameter subtended 3.2 degrees of visual angle) that allowed only one cell of the problem to be viewed at any given time.

### 4.2 Experiment 1

The order of problem information was actively manipulated to quantify the influence of strategic processing on relational reasoning. Specifically, we quantified the boost in performance associated with pure systematic row-wise scanning and the deficit associated with pure haphazard scanning compared to a free-view condition.
4.2.1 Method

Thirty-eight university students with normal or corrected-to-normal vision completed two short-form tests from Raven’s Advanced Progressive Matrices, Set II (Raven et al., 1998) on two separate days approximately a week apart. Half of them completed items 2, 4, 6, 9, 10, 11, 16, 17, 19, 21, 23, 24, 26, 29, 32, 33, and 35 on the first session and 1, 3, 5, 7, 12, 13, 14, 15, 18, 20, 22, 25, 27, 28, 31, 34, and 36 on the second session. The other half completed the same subsets in the opposite order. The instructions followed the Raven APM Manual guidelines for individual test administration (Raven et al., 1998). Participants were paid $6 per hour plus $1 bonus for each correct answer.

The role of strategic processing was evaluated by comparing constrained free-viewing to constrained systematic and haphazard viewing. On the first session all participants completed the test under normal viewing conditions in which the entire item appeared on screen and they could freely fixate the problem information in the order they wished. On the second session participants could only see one cell of the problem at once and the order of viewing was constrained to be systematic (N=13), haphazard (N=12), or free-view (N=13). In the systematic condition the order of problem information was always constrained to row-wise scanning (AOIs 123456789; See Figure 4.1, right). In the haphazard condition the order of problem information for each trial was randomly constrained to one of 22 haphazard scanning patterns (e.g. 492735186) which minimized row-wise and column-wise scanning. In the systematic and haphazard viewing conditions participants transitioned through the problem information using the left/right arrow keys and the 8 possible response alternatives were always viewed systematically across row 1 and 2. The third condition was a control condition where participants could freely control which cell of the problem was viewed using the mouse. For all conditions the outline of the problem matrix
and response area were visible at all times to make it easy for participants to orient to where they were looking in the problem or response areas.

Each trial began with a brief alert sound. A fixation cross appeared on a 21” CRT monitor in a darkened room (Figure 4.1, center). After the participant fixated for 1 s, the Raven problem appeared and remained onscreen until the participant selected a response using the mouse (in the free-view condition) or by pressing enter (in the systematic and haphazard conditions). APM item accuracy, solution time, eye movements, and “think aloud” verbal protocols were collected for each item. Relational performance was assessed based on total number of APM items solved correctly and the total solution time.

4.2.2 Results and Discussion

Prior to analysis of the condition manipulation we verified that pretest performance was consistent with previous APM findings, that the members of each condition were of comparable ability, and that the short-form test sets were also comparable. Trend analyses verified that error-rate linearly increased as a function of trial, $F(1,576) = 150.06, p < .0001, \eta_p^2 = .21$, and solution time increased with a strong linear, $F(1,576) = 183.60, p < .0001, \eta_p^2 = .24$, and mild quadratic component, $F(1,576) = 18.05, p < .0001, \eta_p^2 = .03$, as a function of trial. These findings are consistent with previous studies of APM performance (Bors & Vigneau, 2003; Denney & Heidrich, 1990) and reflect the increasing relational complexity of items as the APM test progresses. A one-way ANOVA with group as a between-subjects factor revealed that there were no significant differences in group APM score, $F(2,34) = 1.49, p = .24$, or solution time, $F(2,34) = .26, p = .77$, at pretest. Finally, two independent samples t-test revealed that performance on the two short-form APM tests were not

1Eye movement and verbal protocols analyses are beyond the scope of this article.
Figure 4.2 RAPM score and solution time during pretest and active strategy manipulation by condition. Left: The number of APM items answered correctly during pretest and for free-view, haphazard, and systematic scanning conditions on session 2. The constrained free-view was not significantly different than pretest. Haphazard scanning produced a significant deficit in APM score and systematic row-wise scanning produced a significant increase in APM score. Error bars represent standard error of the mean. Right: Total APM solution time during pretest and for free-view, haphazard, and systematic scanning conditions on session 2. Significant increases in solution time were observed for the haphazard and systematic viewing conditions relative to pretest. Error bars represent the standard error of the mean.
significantly different at pretest (APM score, $t(35) = 1.36, p = .18$; Solution time, $t(35) = 0.16, p = .87$). Taken together these findings replicated known findings about the APM test and confirmed there were no discernible differences in group ability or short-form test difficulty.

The main question of interest concerning the strength of strategic influence on relational extraction was examined within a mixed ANOVA design with condition as a between-subjects factor (systematic, haphazard, free) and session (pretest, posttest) as a within-subjects factor. The hypothesized influence of the active manipulation of problem information on APM score was observed as a significant condition x session interaction, $F(2, 34) = 5.93, p < .01, \eta^2_p = .26$. Paired one-sided t-tests were used as post-hoc tests to quantify the strength of the increase due to row-wise systematicity, $t(12) = 2.69, p_{adj} < .05, R^2 = .38$, and decrease due to haphazard scanning, $t(10) = 2.03, p_{adj} = .05, R^2 = .30$. No significant change was observed in the free-view condition. These results offer direct support for the hypothesis that the probability of relational insight is directly influenced by the sequence in which information is attended.

In addition to APM score we also examined how processing speed was affected by the active manipulation of scanning strategy. We evaluated solution time using the same mixed ANOVA design with condition as a between-subjects factor (systematic, haphazard, free) and session (pretest, posttest) as a within-subjects factor. The ANOVA revealed only a main effect of session, $F(1, 34) = 14.03, p < .001, \eta^2_p = .29$. This suggests that the active manipulation of problem information that masked all but one cell, slowed processing speed across all three conditions at posttest. Post-hoc

---

2 All post-hoc tests were adjusted for multiple comparisons according to the False Discovery Rate (FDR) correction for multiple comparisons (Benjamini & Hochberg, 1995).
paired one-tailed t-tests revealed significant increases in solution time for the systematic, \( t(12) = 3.12, p_{adj} < .01, R^2 = .45 \), and haphazard conditions, \( t(12) = 4.46, p_{adj} < .001, R^2 = .67 \). The overall increase in solution time may be a result of both the reduced problem information present in each viewing condition and/or the keyboard/mouse problem navigation. It is unclear why the deficits were larger in both the haphazard and the systematic scanning condition compared to the control. This may be a result of more efficient problem navigation using the mouse compared to the keyboard arrow keys.

4.3 Experiment 2

The active manipulation of the sequence of attended problem information in Experiment 1 required a large reduction in the amount of overall information that was actively available to participants. Experiment 2 employed a gaze-contingent viewing paradigm to study the effects of constrained attentional scope.

4.3.1 Method

Forty university students with normal or corrected-to-normal vision completed two short-form tests from Raven’s Advanced Progressive Matrices, Set II (Raven et al., 1998) on two separate days approximately a week apart. Half of them completed items 2, 4, 6, 9, 10, 11, 16, 17, 19, 21, 23, 24, 26, and 29 on the first session and 1, 3, 5, 7, 12, 13, 14, 15, 18, 20, 22, 25, 27, and 28 on the second session. The other half completed the same subsets in the opposite order. The instructions followed the Raven APM Manual guidelines for individual test administration (Raven et al., 1998). Participants were paid $6 per hour plus $1 bonus for each correct answer.

The effect of constrained attentional scope was evaluated by comparing unconstrained viewing at pretest and gaze-contingent viewing at posttest. During the first
session all participants completed the APM test items under normal viewing conditions in which the entire item appeared on screen and they could freely view the entire image. On the second session participants completed each item using a gaze-contingent viewing window that participants controlled by fixating different parts of the screen. The gaze-contingent window was circular and constrained the amount of problem information actively available at any given time to a single cell of the matrix or response area. Just as in Experiment 1, each trial began with a brief alert sound. A fixation cross appeared on a 21” CRT monitor in a darkened room (Figure 4.1, center). After the participant fixated for 1 s, the Raven problem appeared and remained onscreen until the participant selected a response by clicking their chosen answer using the mouse.

4.3.2 Results and Discussion

Prior to analysis of the gaze-contingent manipulation we once again verified that pretest performance was consistent with previous APM findings and that the short-form test sets were comparable. Trend analyses verified that error rate linearly increased as a function of trial, \( F(1, 507) = 76.63, p < .0001, \eta^2_p = .13 \), and solution time increased with a strong linear, \( F(1, 507) = 339.78, p < .0001, \eta^2_p = .40 \), and mild quadratic component, \( F(1, 507) = 37.05, p < .0001, \eta^2_p = .06 \), as a function of trial. These findings are consistent with Experiment 1 and previous studies of APM performance (Bors & Vigneau, 2003; Denney & Heidrich, 1990) and reflect the increasing relational complexity of items as the APM test progresses. Finally, two independent samples t-tests revealed that performance on the two short-form APM tests were not significantly different at pretest (APM score, \( t(38) = 0.33, p = .76 \); Solution time, \( t(38) = 0.53, p = .60 \)). Taken together these findings replicated known findings about
The effect of the gaze-contingent paradigm on relational performance (APM score and solution time) was examined using 2 one-tailed paired t-tests. The results did not support a significant decrease in APM score, $t(39) = 0.76, p = .23$, or increase in solution time, $t(39) = 1.43, p = .08$ in the gaze-contingent viewing condition. This is a surprising finding and suggests that even a large reduction in the amount of information available to brief memory stores such as iconic working memory does not cause significant deficits in overall relational performance.
4.4 General Discussion

In this article we actively manipulated the order in which Raven APM problem information was attended and actively constrained the amount of information available to investigate the roles of strategic processing and visual attention in individual differences in fluid intelligence. Experiment 1 which actively manipulated the presentation order of APM problem information revealed a significant boost in APM score when problem information was viewed systematically and a significant deficit during haphazard presentation. These results experimentally establish the link between individual differences in strategy and relational performance on the Raven APM test. Experiment 2 employed a gaze-contingent viewing paradigm to study the effects of constrained attentional scope. The results did not reveal any significant deficits due to constrained attentional scope.

Previous findings have indicated that information processing strategy is correlated with individual differences in performance on geometric analogy tasks like the Raven APM test (Hayes et al., 2011; Snow, 1980; Bethell-Fox et al., 1984; Carpenter et al., 1990). In all these studies eye movements were used to help decipher how attentional control is deployed over time to infer broader processing strategies. While these earlier studies have indicated that systematic encoding of problem information is beneficial and is more likely to be employed by high Gf individuals, they do not indicate whether this is due to a spurious correlation with fluid ability or a deeper relationship between sequential information intake and the probability of relational insight. By actively constraining the way in which problem information is attended by participants, Experiment 1 provides the first experimental evidence indicating that the order in which Raven APM information is attended directly affects the probability of successful relational extraction.

In Experiment 1 the interaction between processing strategy and relational insight
was evident in APM score, but we also observed an overall increase in solution time across the scanning conditions on session 2. The increase in total solution time in Experiment 1 was most likely driven by the use of the arrow keys and mouse to cycle through each cell of the APM items. This explanation seems plausible because no significant change in solution time was observed in the gaze-contingent paradigm in Experiment 2, which allowed rapid control of the problem cell being viewed using eye movements. It is possible that the slower navigation caused by the arrow and mouse problem navigation obscured processing speed affects of the strategy manipulation. Therefore, one limitation of Experiment 1 is that it did not allow a clear picture of how processing speed may be affected by systematic and haphazard information processing strategies. Moving forward it would be beneficial to find a way to constrain the order of information while still maintaining efficient navigation of problem information. Our results indicate that mouse navigation may be a tenable option as it was the only condition that did not produce a significant effect.

The gaze-contingent viewing paradigm did not produce any significant differences in relational performance. The null result suggests that participants performing matrix reasoning limit their attention to one cell at a time rather than deploying attentional resources across multiple cells. This type of cell-focused attentional deployment also suggests that the strategy manipulation in Experiment 1 (which required reducing the amount of problem information to a single cell) was not confounded by constraining available problem information to a single cell.

One limitation of the gaze-contingent paradigm in Experiment 2 is that it analyzed differences in how attention is deployed only at the group level, collapsing across individual differences. Moving forward it will be important to examine whether there are individual differences in visual attention during relational reasoning that may be masked within this broader effect. For instance previous work has suggested that
high Gf individuals have greater attentional capacity than low Gf individuals (Cowan et al., 2005; Cowan, 2001). Therefore, it is plausible that high Gf individuals may be able to use this greater attentional capacity to deploy attention across multiple cells at once. In future research it would likely be informative to try to isolate individual differences in attentional capacity, attentional scope, and fluid ability by combining latent variable analyses with gaze-contingent active manipulation of RAPM problem information to ascertain whether there are individual differences in how visual attention is deployed during relational reasoning.

Using active experimental manipulation during relational reasoning on the Raven APM test provided novel insights into how strategy and visual attention contribute to relational performance. Experiment 1 firmly established a causal relationship between strategy and the probability of relational insight. The gaze-contingent paradigm in Experiment 2 revealed a surprising finding that constraining attention to a single cell in the APM problem matrix did not affect relational reasoning performance. This aggregate analysis suggests that during matrix reasoning participants deploy their attention cell-by-cell rather than across multiple cells. Further research combining latent variable analyses and active manipulation during fluid reasoning will likely prove useful in identifying whether individual differences exist between high and low Gf individuals that may be masked by this broader effect.
CHAPTER 5

Mapping and correcting the influence of gaze position on pupil size measurement

5.1 Introduction

The human pupillary response correlates with a wide range of important cognitive variables including mental workload (e.g., Hess & Polt, 1964; Kahneman, 1973; Klingner, Tversky, & Hanrahan, 2011), emotional valence (Partala & Surakka, 2003), attention (Beatty, 1982a), working memory (Beatty & Kahneman, 1966), arousal (Murphy, Robertson, Balsters, & O’Connell, 2011), decision-making (Einhäuser et al., 2010), surprise (Preuschoff, Marius, & Einhäuser, 2011), and uncertainty (Nassar et al., 2012). While the underlying mechanisms that drive cognitive pupillary effects are still an active topic of inquiry, recent data suggests that the pupillary response may reflect noradreneric activity in the brain (Murphy et al., 2014; Aston-Jones & Cohen, 2005; Koss, 1986; Samuels & Szabadi, 2008a). As a result of these recent developments, pupil size is increasingly being used by cognitive scientists as an important measure of cognitive processing.

Pupil data can be collected inexpensively and non-invasively by most modern

The text of this chapter [5] is taken from Hayes & Petrov currently under review in Behavior Research Methods
image-based eye-tracking systems, but researchers often underestimate the methodological challenges in producing well-controlled studies that allow for cognitive interpretations of the pupillary response (Holmqvist et al., 2011). Most researchers are aware of the most serious confound—the pupillary light reflex. The pupil diameter depends first and foremost on the luminance of the stimulus and the ambient illuminance of the experimental room. Changes in either of these variables can produce strong light reflexes that modulate pupil size by as much as 50% (Miller & Newman, 2005; Loewenfeld, 1993). Cognitive researchers control for this confound by using isoluminant stimuli and maintaining consistent room illumination. A lesser known yet equally problematic confound is pupil foreshortening error (PFE). Changes in gaze position produce foreshortening of the pupillary image because the eye tracking camera is fixed but the eye rotates. Specifically, as the eye rotates away from the camera, the pupil image becomes more elliptical and shrinks in apparent area by as much as 10%. This is a large margin compared to the magnitude of the cognitive effects, which rarely exceed 5% change in pupil size. Despite its seriousness, the PFE is not corrected in many commonly used remote eye-tracking systems (e.g., Tobii, Tobii Technology AB, 2010, and EyeLink, SR Research, 2006). All too often, the PFE is not controlled by the experimental design either (Brisson et al., 2013). The combination of these factors has led to an increasing number of published studies with inadequate control for PFE.

There are currently two different methods aimed at circumventing PFE using experimental design. The most conservative method—and the one recommended by most commercial eye-tracking providers—is to have participants maintain constant fixation throughout each trial (e.g., SR Research, 2006, p. 98). The constant fixation method is implemented by defining a small fixation boundary area and verifying that the eye position remains within this area throughout the pretrial fixation and stimulus
presentation periods. This minimizes PFE by keeping the optical axis of the eye at a fixed angle from the camera. However, the constant fixation method has two serious drawbacks. First, it limits the types of tasks that can be studied to those with a constant fixation. Unfortunately, this rules out studies of important topics such as visual search or reading. Second, it is hard for participants to consistently maintain fixation for long periods of time. This results in a high number of invalid trials that increases as the stimulus duration increases. The other PFE-mitigation method that is sometimes recommended by eye-tracking providers (SR Research, 2006, p. 98) is stimulus-position counterbalancing. In these designs, there are several positions where the stimulus can appear, but the number of presentations is counterbalanced across trials. This method still does not allow the study of tasks that require free viewing of the stimulus. Neither does it allow for comparisons of physical pupil size among different gaze locations because PFE varies across the visual field (Gagl, Hawelka, & Hutzler, 2011). In sum, though useful, both methods have serious limitations.

A completely different approach to the problem is to measure PFE and then correct the pupillometric data prior to analysis (Gagl et al., 2011; Brisson et al., 2013). Gagl and colleagues (2011) recently pioneered this approach with the aid of an artificial eye model with a fixed pupil size. The artificial eye was moved across a single horizontal scan line, emulating a sentence reading task. The PFE could thus be measured explicitly and a correction formula was developed. The horizontal extent of the sentence stimuli spanned from $-17^\circ$ to $+9^\circ$ from the screen center. This resulted in systematic pupillometric error from $+5\%$ to $-13\%$, respectively, relative to the true pupil area of the artificial eye. These data allowed the development of a mathematical model of the PFE. This model was used to correct the empirical data from a sentence reading task and a “Z-string” reading task in which participants scanned words composed entirely of the letter Z. When the measurement error was
Figure 5.1 *Photograph of the artificial eye, ocular socket, and pupil calibration devices.* Each artificial eye was measured in a constant left-eye position within the chin- and forehead-rest. The pupil calibration apparatus was not present during artificial-eye data collection as shown, but rather replaced the ocular socket during the pupil calibration procedure.
corrected, previously reported discrepancies between the pupillary response during word recognition and sentence reading tasks were revealed to be artifacts of PFE (Gagl et al., 2011). This pioneering study illustrates the importance of accounting for PFE and how it can lead to incorrect cognitive interpretations if not properly controlled or corrected. One major shortcoming of this study is that it only mapped 11 points along a single horizontal scan line. The resulting correction formula is applicable to this special case only.

Brisson and colleagues (2013) used an object pursuit task to characterize the PFE of three popular eye-tracking systems (Tobii X120, Tobii T120, and EyeLink 1000) across both horizontal and vertical changes in gaze position. Forty-four human participants tracked a circle moving counterclockwise across the display in an elliptical pattern that covered up to 22° of horizontal visual angle and up to 14° of vertical visual angle (depending on the system). The results showed substantial measurement errors in pupil diameter as a function of gaze position in all three systems. A linear regression model using X and Y gaze coordinates as predictors was able to explain between 9% and 20% of the pupil variance depending on the system and the maximum visual angle. While these data provided important new insights into the extent of PFE with human participants across multiple eye-tracking systems, the regression-based correction procedure only accounted for a small amount of the total PFE. Furthermore, even low-effort tasks such as object tracking require attentional mechanisms and cognitive effort that induce pupillary variability (Beatty, 1982a, 1982b; Kahneman, 1973) that complicates the estimation of the foreshortening error.

In the present article we systematically map and correct pupil foreshortening error across the full display using a set of artificial eyes with fixed, known pupil diameters. We report two studies. The first formally verified the ratio scale between the “arbitrary units” of the EyeLink 1000 and physical units. The second study systematically
maps PFE and develops both a parameter-free and a parameterized geometric model correction that virtually eliminates PFE. Three artificial eyes were measured across the full extent of a 21” monitor in each of three separate experimental layouts varying the relative distances between the camera, monitor, and eye. The measurements revealed large distortions in recorded pupil diameter as a function of gaze position. A simple geometric model that estimates pupil foreshortening as the cosine of the angle between the eye-to-camera axis and the eye-to-stimulus axis produced excellent fits to the data (cf. Atchison & Smith, 2000; Spring & Stiles, 1948). The model reduced the root mean squared error of pupil measurements by 82.5% when the model parameters were pre-set to the physical dimensions of the experimental layout, and by 97.5% when they were optimized to fit the empirical error surface. Thus, our calibration
data and correction procedure allow for an unprecedented reduction in PFE without sacrificing data quality or experimental flexibility in the process.

5.2 Experiment 1

The *EyeLink 1000 User Manual* (SR Research, 2006, p. 98) states that, “The pupil size data is not calibrated, and the units of pupil measurement will vary with subject setup. Pupil size is an integer number, in arbitrary units. . . . Pupil size measurements are affected by up to 10% by pupil position, due to the optical distortion of the cornea of the eye, and camera-related factors.”

This statement is very vague. It does not even specify the type of scale for these “arbitrary units”. A measurement procedure establishes a correspondence between a set of numbers and a set of objects with respect to some attribute of interest—pupil diameter in our case. It matters greatly what types of relationships are preserved by this mapping (Torgerson, 1958). An *ordinal scale* only preserves comparative relations (<, =, >). An *interval scale* preserves distances but not ratios. Thus, it is meaningful to calculate means, differences, and standard deviations among interval-scale measurements, but they do not warrant inferences of the form, “the pupil diameter in condition A is 10% greater than that in condition B.” The latter kind of statement asserts a multiplicative relationship ($A = 1.1 \times B$) that is warranted only for a *ratio scale*—that is, for an interval scale with a *true origin*. Ratio scales map the number zero to the (possibly hypothetical) object that altogether lacks the attribute in question.

Ideally we want to measure pupil diameter on a ratio scale with standard units such as millimeters. The next best method is to measure it on a ratio scale whose units are proportional to millimeters, although the coefficient of proportionality is not specified. This is one possible interpretation of the phrase “arbitrary units.”
The coefficient of proportionality may vary with experimental setup, as the *Manual* warns, but as long all measurements within a given setup form a ratio scale, they still support multiplicative operations. This is important because, as we show below, the geometric foreshortening law is multiplicative: The foreshortened diameter equals the true diameter *times* the cosine of a certain angle.

The *Manual* provides very little information about how the pupillometric data are acquired and processed by the proprietary software on the host computer. This raises the disconcerting possibility that the “arbitrary units” may only form an interval scale. This could occur, for instance, if the software added an unspecified constant before writing the numbers to the data file. This would render invalid all multiplicative operations with these measurements. In particular, it would undermine the common practice in the literature to report the pupillometric data in terms of percent change from baseline.

Finally, the “arbitrary units” may form merely an ordinal scale. This could happen, for instance, due to “the optical distortion of the cornea . . . and camera-related factors.” In this case, the information content of the pupillometric data would be very low and they should be analyzed using ordinal statistical methods (Agresti, 1984).

The purpose of this preliminary experiment is to collect calibration data to characterize the type of scale used for EyeLink 1000 pupillometry. The results indicate that the “arbitrary units” form a ratio scale with a layout-dependent coefficient of proportionality to millimeters.

### 5.2.1 Method

We used an EyeLink 1000 desktop eye tracker (SR Research, 2006) at a sampling rate of 1000 Hz. The tracking mode was set to “pupil only,” the pupil threshold parameter to 110, and the eye-to-track parameter to “Left.” Pupil area was measured using
centroid mode throughout the study. Centroid mode computes pupil area using a center-of-mass algorithm that identifies the number of black pixels in the thresholded pupil and its center on the camera image. The User Manual recommends the center-of-mass algorithm over the ellipse fitting algorithm because “it has very low noise” (SR Research, 2006, p. 71).

A series of 11 black discs with diameters ranging from 2 mm to 7 mm in 0.5 mm increments were printed on white paper using a laser printer. The accuracy of the printed diameter was verified with Neiko digital calipers. An apparatus consisting of a flat piece of wood (157 mm x 88 mm x 36 mm) and an opaque white piece of plastic (50 mm x 95 mm) with a window (13 mm x 13 mm) served to present the printed pupils at a constant position (see Figure 5.2a). The 7 pupils were printed on a single tape that could be placed underneath the plastic pouch and translated horizontally.

We explored three separate geometric layouts—Near, Medium, and Far—that varied the relative positions of the camera, artificial eye, and monitor as specified in Figure 5.3, panels a, b, and c, respectively. At each layout we made pupillometric recordings with each printed pupil for 5 seconds at 1000 Hz, thereby obtaining 5000 pupil-area samples. The area data were converted to linear (diameter) data by taking square roots, and then averaged across the 5000 samples. This procedure produced a calibration data set of 33 pupil diameter measurements (3 layouts × 11 pupils).

5.2.2 Results and Discussion

Figure 5.4 plots the calibration data in EyeLink’s “arbitrary units” against the the true diameter of the physical pupil in mm. The results clearly indicate that the arbitrary units form a ratio scale within each layout, but the coefficient of proportionality to the true pupil diameter varies across layouts. Specifically, this coefficient was $k_n = 10.07$ for Near, $k_m = 9.65$ for Medium, and $k_f = 8.43$ for the Far layout in
Figure 5.3 *Experimental Layouts and Geometric Model.* Three separate geometric layouts (panels a–c) varied the relative positions of the camera, artificial eye, and monitor. Each layout is diagrammed from three vantage points (top, side, and front). All distances are in millimeters. Pupil calibration (Experiment 1) and pupil foreshortening (Experiment 2) were measured for each layout. The geometric model (d) estimates the foreshortening of the pupil area as a function of the cosine of the angle $\theta$ between the eye-to-camera axis $OC$ and the eye-to-target axis $OT$. The origin $O$ of the coordinate system is at the center of the artificial pupil. The X axis is horizontal and parallel to the bottom edge of the screen, growing rightward. The Y axis is vertical, growing upward. The Z axis is perpendicular to the screen, growing inward.
Figure 5.4 *Ratio scale between arbitrary units and physical units* The pupil-diameter data recorded by EyeLink 1000 in arbitrary units are proportional to the true diameter of the physical pupil, but the coefficient of proportionality depends on the distance between the camera and the artificial eye.

Figure 5.3. Each of these linear regressions accounted for essentially all the variance ($R^2 > .999$) of the 11 measurements within a given layout. Importantly, adding a free intercept parameter to the regression equation did not significantly improve the fits, and the intercept estimates were negligible ($< 0.5$ AU). This indicates that 0 arbitrary units always correspond to 0 mm regardless of layout, thereby satisfying the true-origin requirement for ratio scales.

Furthermore, the slope coefficients $k$ were themselves approximately inversely proportional to the eye-to-camera distances, which were $L_n = 584$, $L_m = 610$, and
$L_f = 698$ mm for the three layouts. To verify this inverse proportionality, we calculated products of the form $p_i = \alpha k_i L_i$, where $i \in \{n, m, f\}$ and $\alpha \approx 1.70 \times 10^{-4}$ is a re-scaling parameter determined in the next paragraphs. If each slope coefficient is inversely proportional to the eye-to-camera distance regardless of layout, then $p_i$ should equal unity for any $i$. In our data, we obtain $p_n = 0.9999$, $p_m = 1.0002$, and $p_f = 0.9998$.

This suggests that the “arbitrary units” are not really units of length but of visual angle. Apparently, the EyeLink 1000 in centroid mode reports the angular area subtended by the pupil. A straightforward algorithm for estimating this angular area—and probably the algorithm implemented by the EyeLink software—is to count the “pupil” pixels in the image of the eye. This interpretation is consistent with the statement in the User Manual that the pupil area is recorded in “scaled image pixels” (p. 17). Taking the square root transforms the angular area into a linear visual angle $\phi$, which is approximately proportional to the ratio of the true pupil diameter $d$ and the viewing distance: $\phi \propto d/L$.

On the basis of this invariant relationship, Equation 5.2.1 provides a straightforward formula for converting “arbitrary units” to mm at any viewing distance:

$$d \approx \alpha L \phi$$

(5.2.1)

In this equation, the angle $\phi$ subtended by the pupil is reported by the EyeLink 1000 in “arbitrary units,” $L$ denotes the eye-to-camera distance in mm, and the re-scaling parameter $\alpha$ is in radians per arbitrary unit. We estimate $\alpha$ from the relationship $d_j/(\phi_{ij} L_i) = \text{const}$ across the 11 pupils $j$ and 3 layouts $i$. This ratio has a mean of

---

1In Figure 5.3, these are the hypotenuses of the triangle with vertical side $C_y = 310$ mm giving the height of the eye relative to the camera and with horizontal side $C_z = 495, 525$, and $625$ mm, respectively.

2We use the paraxial approximation $\tan \phi \approx \sin \phi \approx \phi$ for small angles measured in radians (Atchison & Smith, 2000).
\[ \alpha = 1.70 \times 10^{-4} \text{rad/AU} \approx 35.1 \text{arcsec/AU} \text{ in our data set (SD = 7.20 \times 10^{-7} \text{rad/AU} \approx 0.149 \text{arcsec/AU}).} \]  With just one free parameter, Equation 5.2.1 accounts for 99.99\% of the variance of the 33 calibration measurements.

In conclusion, Experiment 1 established that EyeLink’s “arbitrary units” form a ratio scale with coefficient \( \alpha L \) depending on the eye-to-camera distance \( L \) for each fixed layout. Also, the pupil-size data seem to track the visual angle subtended by the pupil. Consequently, the units can be converted across layouts according to Equation 5.2.1. These results were obtained with very simple artificial pupils printed on a flat surface. The next experiment investigates the pupil foreshortening error caused by the rotation of spherical artificial eyes relative to the camera.

### 5.3 Experiment 2

When an eye is photographed from an oblique angle, the image of the pupil becomes elliptical and the apparent pupil area decreases (e.g., Jay, 1962; Spring & Stiles, 1948). The resulting pupil foreshortening error (PFE) depends on the viewing angle. In a typical eye tracking setup, the camera is stationary relative to the screen but the eye rotates in its socket, thereby varying the angle between the optical axis of the camera and the plane of the pupil. We constructed spherical artificial eyes that could rotate in artificial sockets and could be pointed to “fixate” arbitrary points on the screen. The purpose of Experiment 2 was to map the relationship between EyeLink’s pupillometric data and the screen coordinates of a grid of fixation points. The results indicated that the measurements were contaminated by substantial PFE, but the magnitude of the error depended systematically on the cosine of the angle between the eye-to-camera axis and the eye-to-stimulus axis. Because the PFE was systematic, it could be corrected very well by a simple geometric model.
5.3.1 Method

Apparatus

All EyeLink 1000 settings were the same as those in Experiment 1, with the exception of the pupil threshold parameter. The latter was lowered from 110 to 60 as the spherical artificial eyes with pupillary wells required a lower threshold than the flat pupils printed on paper. These settings provided a reliable thresholded pupil signal across all measured conditions.

Artificial Eye Model

Three artificial eyes were manufactured from three solid wooden spheres, each 31 mm in diameter. The production of each eye began by drilling a 0.79 mm pilot hole all the way through the center of the wooden ball using a Skil 3320 10” benchtop drill press and DeWalt titanium pilot drill bits. Then a well was drilled 20 mm deep into each eye to form the artificial pupil (see Figure 5.2b, c). The three eyes had well diameters of 3.17 mm, 4.76 mm, and 7.14 mm, respectively. The inside of each well was painted black and the rest of the eye was painted white using tempera washable paint. Finally, a pressure-switch laser sight (<5 mW) was collimated within a 100-mm section of PVC pipe (25 mm diameter) that was attached to a second 30-mm section of PVC pipe (15 mm diameter) that was firmly cemented to each artificial eye using epoxy cement (Figure 5.2c). The laser beam originated approximately 75 mm behind the eyeball and traveled to the screen through the .79-mm hole. The pressure switch allowed us to (de)activate the laser without disturbing the eye orientation in the socket.

The artificial eye was kept at a constant position within the headrest using a simple ocular socket mechanism (see Figure 5.1). This mechanism was composed of 4 pieces of wood (sides 131 x 13 x 28 mm; top and bottom 13 x 63 x 38 mm), 2 steel mending
plates (140 x 0.9 x 13 mm), 2 threaded steel rods (152 mm long), 4 wing nuts, and 2 rubber grommets. Two holes were drilled through the wooden sides for the threaded rods—one 14 mm from the top and one 26 mm from the bottom. Two circular recessed holes (25 mm diameter) were drilled 4 mm deep into the inside of each wooden side piece (centered horizontally and vertically) to hold the rubber grommets firmly in place. The ocular socket was then assembled as shown in Figure 5.2d. The threaded rods were placed through each side piece with a mending plate on each outside edge held in place by the four wing nuts. The wing nuts could then be tightened down to firmly hold the current artificial eye in place and loosened to remove and replace the eye when needed. The elasticity of the rubber grommets allowed us to manipulate the elevation and azimuth of the artificial eye within the socket, while at the same time holding firmly when the desired orientation was reached.

The standard EyeLink 1000 nine-point calibration and validation procedures were used to quantify the accuracy of the artificial eye and socket apparatus. The nine-point validation data indicated high gaze-position accuracy across the display (average mean error M=0.18, SD=0.04; maximum error M=0.26, SD=0.04 degrees of visual angle). The validation data verified that the laser pointers were collimated accurately and the ocular socket kept the artificial eyes in a stable orientation throughout the procedure.

**Gaze Position Map and Procedure**

Pupil size measurements were collected with each artificial eye on a rectangular grid of fixation targets for each experimental layout. For the Near and Far layouts, we used $16 \times 12 = 192$ targets spaced 64 pixels apart (see Figure 5.2e top); for the Medium layout, we used $8 \times 6 = 48$ targets spaced 128 pixels apart (at screen resolution $1024 \times 768$).
On each trial, we rotated the eye until the pointer beam illuminated the appropriate target on the screen. A box shielded the eye tracking camera as a precaution against accidental exposure to the beam. As in Experiment 1, pupil-area measurements were recorded for 5 s at 1000 Hz, converted to linear units by taking square roots, and averaged across the 5000 samples. This produced one pupil-diameter datum per trial.

**Geometric Model**

The results from Experiment 1 suggested that the EyeLink 1000 (operating in centroid mode) reports the *angular* area subtended by the pupil. Based on simple geometric principles, Equation 5.2.1 accounted extremely well for the pupillometric data when the distance between the camera and the eye was varied. Our working hypothesis is that the same principles can account for the pupillometric data when the angle between the camera axis and the eye axis is varied. In this section, we develop a geometric model that formalizes this hypothesis and provides explicit formulas for correcting the pupil foreshortening error.

The key term in the model is the *oblique angle* \( \theta \) between the eye-to-camera axis and the eye-to-stimulus axis (Figure 5.3d). To simplify the trigonometric calculations, the model assumes the camera is pointed directly at the eye and thus the eye-to-camera axis coincides with the optical axis of the camera. Let us consider first the baseline configuration in which the eye too is pointed directly at the camera. Then the eye axis coincides with the camera axis and \( \theta = 0 \). Let \( A_0 \) denote the angular area subtended by the pupil as viewed from the camera in the baseline configuration.

Our goal is to quantify the foreshortening effect, whereby the apparent angular area \( A(x, y) \) diminishes when the eye rotates away from the camera to fixate a target with coordinates \( x \) and \( y \) on the screen. The derivation proceeds in two steps: First,
we argue that the foreshortening is multiplicative and depends on the oblique angle $\theta$ according to Equation 5.3.1. Second, we express $\cos \theta$ as a function of $x$ and $y$ using simple vector calculus.

$$A(x, y) = A(\theta(x, y)) = A_0 \cos \theta(x, y) \quad (5.3.1)$$

We consider a circular pupil for concreteness, although the mathematical derivation generalizes to arbitrary planar shapes. When a circular disk is viewed from an oblique angle, it is projected into an ellipse and its apparent area decreases (Gagl et al., 2011; Jay, 1962; Mathur, Gehrmann, & Atchison, 2013; Spring & Stiles, 1948). Importantly, the projected pupil becomes narrower in the direction of view (the tangential section) but remains unchanged in the perpendicular direction (the saggital section, Equation 3.2 in Atchison & Smith, 2000). This is why the multiplier in Equation 5.3.1 above is $\cos \theta$ rather than $\cos^2 \theta$. To simplify the analysis, we have assumed that the eyeball diameter is negligible relative to the distance to the camera and, consequently, the rotational center of the eyeball lies approximately in the pupil plane. This seems a reasonable assumption for desktop trackers such as the one used in this article, but probably becomes very crude for head-mounted trackers. Taking the eyeball geometry into account does not change any of the principles here, but complicates the trigonometry. Another assumption that is implicit in Equation 5.3.1 is that the camera has negligible optical aberrations near the center of its field of view.

The second step of the derivation is to express $\cos \theta$ as a function of the target coordinates $x$ and $y$. It is convenient to work in a coordinate system with an origin $O$ at the pupil center and axes as described in the caption of Figure 5.3d. The camera lens is at point $C$ with coordinates $C_x$, $C_y$, and $C_z$, which are parameters of the physical layout (Figure 5.3). The other parameters of the layout are the coordinates $S_x$, $S_y$, and $S_z$ of the upper left-hand corner of the screen. Then a fixation target
T with screen coordinates \( x \) and \( y \) mm has coordinates \( T_x = x - S_x \), \( T_y = y - S_y \), and \( T_z = S_z \) in the eye-centered system. In this notation, the oblique angle \( \theta \) is the angle \( \text{COT} \) in Figure 5.3d and its cosine can be calculated via the dot product of the vectors \( \text{OC} \) and \( \text{OT} \):

\[
\cos \theta(x, y) = \frac{C_x T_x + C_y T_y + C_z T_z}{\sqrt{C_x^2 + C_y^2 + C_z^2 \sqrt{T_x^2 + T_y^2 + T_z^2}}} \quad (5.3.2)
\]

For ease of comparison to the physical pupil diameter of the artificial eye models, the angular area \( A \) is converted to arc length \( \phi \) using the relationship \( \phi = 2\sqrt{A/\pi} \).

Taking square roots of both sides in Equation 5.3.1 and rearranging leads to Equation 5.3.3, in which \( \phi_0 \) denotes the angle subtended by the pupil diameter in the baseline configuration and \( \phi(x, y) \) denotes the apparent angle when the artificial eye points at a target with screen coordinates \( x \) and \( y \).

\[
\phi_0 = \sqrt{\frac{\phi(x, y)}{\cos \theta(x, y)}} = \text{const} \quad \text{for all} \quad x, y. \quad (5.3.3)
\]

The true pupil diameter \( d \) can be determined by pointing the artificial eye directly at the camera (or asking the human participant to look at the camera), obtaining EyeLink measurements of the subtended angle \( \phi_0 \) in “arbitrary units,” and converting them to millimeters according to Equation 5.2.1. The pupil foreshortening error \( PFE(x, y) = \phi(x, y)/\phi_0 \) at fixation point \((x, y)\) is \( \sqrt{\cos \theta(x, y)} \). For an artificial eye with a constant pupil, the geometric model predicts that the ratios in Equation 5.3.3 will be invariant for all points \((x, y)\). This prediction is tested below.

Two versions of the geometric model were compared: a parameter-free model and an optimized model. The parameter-free model simply calculated \( \cos \theta \) according to Equation 5.3.2 using the physical measurements taken directly from each experimental layout (Figure 5.3). The relevant measurements are the camera lens coordinates \((C_x, C_y, C_z)\) and the coordinates \((S_x, S_y, S_z)\) of the upper left-hand corner of the screen for the respective layout. The parameter-free model is an idealization that
does not take into account the error of the layout measurements or the optics of the camera lens itself, which can alter the effective geometric layout. In order to account for these additional sources of error, an optimized model was also considered. The optimized model fit 5 parameters $C_x$, $C_y$, $S_x$, $S_y$, and $S_z$; the $C_z$ parameter was fixed to the respective physical layout measurement. The optimized model was fit to the pupillometric data separately for each layout using an unconstrained Nelder-Mead optimization routine that minimized the root mean squared error. The initial values for the iterative parameter search were set to the physical layout measurements.

5.3.2 Results and Discussion

Pupil measurements for the 3 spherical artificial pupil diameters (3.17 mm, 4.76 mm, 7.14 mm) and 3 experimental layouts (Near, Medium, Far) resulted in a total of 9 empirical maps. Figure 5.5 shows the measured pupil diameter as a function of visual angle for each unique map. Pupil diameter is expressed on a common scale across all 9 maps as the deviation from the map-specific geometric mean. Despite the fixed pupil diameter of the artificial eye models, the average range in deviation from the geometric mean for the Near, Medium, and Far Layouts were 14.4%, 10.1%, and 8.4% respectively (see Table 5.1). Pupil diameter systematically decreased as the eye rotated further away from the camera lens—producing larger variability in pupil diameter as viewing distance decreased and visual angles to the target grid points on screen increased. These results establish the existence of large PFE as a result of changes in gaze position for each measured combination of pupil diameter and experimental layout.

PFE was invariant across the 3 different artificial eye diameters within a given experimental layout. Principal-component analysis (PCA, Everitt & Dunn, 2001) was performed to estimate the intra-map similarity across the 3 pupil diameter maps
<table>
<thead>
<tr>
<th></th>
<th>RMSE</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Near Layout</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pupil Diameter 3.17</td>
<td>0.034</td>
<td>0.916</td>
<td>1.058</td>
</tr>
<tr>
<td>Pupil Diameter 4.76</td>
<td>0.032</td>
<td>0.918</td>
<td>1.053</td>
</tr>
<tr>
<td>Pupil Diameter 7.14</td>
<td>0.038</td>
<td>0.905</td>
<td>1.060</td>
</tr>
<tr>
<td><strong>Medium Layout</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pupil Diameter 3.17</td>
<td>0.025</td>
<td>0.940</td>
<td>1.044</td>
</tr>
<tr>
<td>Pupil Diameter 4.76</td>
<td>0.026</td>
<td>0.938</td>
<td>1.043</td>
</tr>
<tr>
<td>Pupil Diameter 7.14</td>
<td>0.024</td>
<td>0.944</td>
<td>1.037</td>
</tr>
<tr>
<td><strong>Far Layout</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pupil Diameter 3.17</td>
<td>0.022</td>
<td>0.945</td>
<td>1.039</td>
</tr>
<tr>
<td>Pupil Diameter 4.76</td>
<td>0.019</td>
<td>0.951</td>
<td>1.030</td>
</tr>
<tr>
<td>Pupil Diameter 7.14</td>
<td>0.019</td>
<td>0.951</td>
<td>1.030</td>
</tr>
</tbody>
</table>

Table 5.1 Descriptive statistics of the deviation in pupil diameter from the geometric mean for the 9 empirical maps (3 experimental layouts x 3 pupil sizes). RMSE = root mean squared error.

within the Near, Medium, and Far Layouts. The first principal component accounted for nearly all the variance across the three pupil diameter maps (Near 99.8%; Medium 98.9%; Far 98.9%). The invariance across changes in pupil diameter is a critical finding as it establishes that the PFE surface is not affected by the changes in pupil diameter that occur when measuring a dynamic, biological eye. Due to the invariance across pupil diameter the maps were collapsed by taking the geometric mean across the 3 different pupil diameter maps and normalizing relative to the geometric mean within a given experimental layout, resulting in one aggregate map for each experimental layout. The remaining results focus on modeling and correcting PFE in the aggregate layout maps.

We evaluated the validity of the geometric model in Equations 5.3.2 and 5.3.3, as
Figure 5.5 Deviation in pupil diameter by pupil size and layout. Each panel shows the deviation in measured pupil diameter from the geometric mean for a unique combination of pupil size (3.17 mm, 4.76 mm, 7.14 mm) and experimental layout (Near, Medium, Far). Large deviations in measured pupil diameter were observed across each map despite the fixed pupil of each artificial eye. Deviations in measured pupil size grew as layout distance became shorter and visual angles increased. Deviations were relatively invariant across pupil diameter within a given experimental layout. Gaze positions are in degrees of visual angle.
Figure 5.6 Aggregate empirical maps, parameter-free geometric model multipliers, and parameter-free correction. Each panel shows the deviation in pupil diameter from the geometric mean. The first column of panels shows the aggregate empirical maps collapsed across pupil size for each experimental layout. The second column shows the corrective multipliers produced by the parameter-free geometric model. The third column shows the corrected data produced by dividing the aggregate empirical data by its respective corrective multiplier. The parameter-free geometric correction reduced the within-map RMSE to 18.7%, 15.8%, and 18.2% of its original value for the Near, Medium, and Far layout. Gaze positions are in degrees of visual angle.
well as its ability to correct the PFE and recover the true pupil diameter $\phi_0$. As a result of the normalization and aggregation described above, $\phi_0 = 1 = \text{const}$ for each map and all residual variation is due to foreshortening error. The parameter-free model reduced the root mean squared error of the aggregated measurements by an average of 82.5%. The correction is applied according to Equation 5.3.3 by dividing the measured pupil diameter at each grid target location by the square root of the cosine of its corresponding angle. Figure 5.6 shows the empirical aggregate layout maps before correction, the geometric model corrective multipliers, and the corrected maps. There was a strong correspondence between the 3 aggregate maps and their respective multipliers (Near $R^2=.99$, Medium $R^2=.98$, Far $R^2=.98$). The parameter-free geometric correction greatly reduced the RMSE to 18.7%, 15.8%, and 18.2% of its original value for the Near, Medium, and Far layout. Table 5.2 shows the descriptive statistics before and after applying the parameter-free correction.

While the parameter-free geometric model reduced the RMSE in pupil diameter substantially, it left a small amount of residual error that was systematic and thus could be corrected further. Specifically, close inspection of the third column in Figure 5.6 suggests that the correction generated by the parameter-free model was not strong enough to fully correct the “tilt” of the uncorrected maps in the first column. Thus, even after the parameter-free correction, the pupil diameter remained slightly underestimated for fixation targets near the left edge of the screen and slightly overestimated for targets near the right edge.

To evaluate the extent to which this residual error can be corrected still further, the parameters of the geometric model were optimized to fit each aggregate map. Figure 5.7 shows the three empirical maps, the optimized corrective multipliers, and the optimized correction maps. The RMSE was reduced to 2.0%, 2.5% and 3.2% of its original value for the Near, Medium, and Far layout, respectively, using the
optimized parameters listed in Table 5.3. As can be seen in Table 5.2, the root mean squared error was less than 0.1% across the three optimized correction maps, which is well within the uncertainty of measurement of a stationary pupil across trials. This indicates that the PFE can be eliminated entirely for practical purposes, at least for the artificial eyes considered here, provided enough calibration measurements are available to constrain the parameters of the geometric model.

It is instructive to compare the optimized parameter values in Table 5.3 to their physical counterparts. The distance $S_z$ from the eye to the screen was constrained well by the data but there were surprisingly large deviations in the other four parameters. The optimized values of $C_x$ and $S_x$ would be veridical if the artificial eye were located $\approx 60$ mm closer to the left edge of the screen (on average across the three layouts). The optimized values of $C_y$ and $S_y$ would be veridical if the eye were located $\approx 90$ mm lower. The overall effect was to produce greater variation in the oblique angle $\theta$ in Equation 5.3.2 across each map, which in turn produced stronger correction multipliers. The within-map RMSE of the optimized model multipliers were larger than their parameter-free counterparts and better matched the RMSE of the calibration data (Table 5.2). We speculate that the optimized parameters captured the magnification effect of the optics inside the eye-tracking camera. Equations 5.3.2 and 5.3.3 essentially model a pinhole camera, whereas the real device includes lenses on the optical path. This is a topic for further research. The theoretical significance of the optimized model fit is that the PFE surfaces were smooth and could be parameterized with a few numbers that could be estimated from calibration data.

5.4 General Discussion

This article reports two experiments that investigate and calibrate the pupillometric measurements of a tabletop EyeLink 1000 eye tracker (SR Research, 2006) using
Figure 5.7 *Aggregate empirical maps, optimized geometric model multipliers, and optimized correction* Each panel shows the deviation in pupil diameter from the geometric mean. The first column of panels shows the aggregate empirical maps collapsed across pupil size for each experimental layout. The second column shows the corrective multipliers produced the geometric model with optimized parameters. The third column shows the corrected data produced by dividing the aggregate empirical data by its respective corrective multiplier. The optimized geometric correction virtually eliminated pupil foreshortening error, reducing the within-map RMSE to 2.0%, 2.5%, and 3.2% of its original value for the Near, Medium, and Far layout. Gaze positions are in degrees of visual angle.
artificial eyes with known pupil diameter. Experiment 1 established that EyeLink’s “arbitrary units” form a ratio scale with a true zero. This result justifies the common practice in the field to express relative pupil diameters as dimensionless ratios—e.g., percent change from baseline. When absolute measurements are needed, they can be calculated for a given arrangement of the eye tracking camera, headrest, and monitor: The “arbitrary units” are proportional to millimeters but the coefficient of proportionality depends on the eye-to-camera distance (Equation 5.2.1). Furthermore, the results of Experiment 1 strongly suggest that EyeLink’s pupil-size data are expressed not in units of linear length but of the angle subtended by the pupil as viewed from the camera. These units can be converted across layouts according to Equation 5.2.1.

Experiment 2 mapped the pupil foreshortening error (PFE) when the artificial eye rotated in its socket to fixate a grid of target points on the screen. Data were collected across 3 experimental layouts with spherical artificial eyes with 3 pupil diameters. The 9 resulting maps showed large PFE that increased as a monotonic function of the oblique angle between the eye-to-camera axis and the eye-to-target axis. The results supported three major conclusions as follows: First, the relative PFE maps were highly replicable across different pupil sizes, as evident in the high ($R^2 > .98$) correlation among maps collected with different artificial eyes in a given experimental layout. This suggests that the relative PFE varies only as a function of the orientation of the eye with respect to the camera but does not depend on the pupil diameter. In other words, the eye-tracking hardware and software operate linearly with respect to pupil diameter. Second, a parameter-free geometric model reduced the error to <18% of its uncorrected value. This suggests that the PFE is largely due to geometric foreshortening of the type captured by Equations 5.3.2 and 5.3.3. Note that the foreshortening follows a multiplicative law (Equation 5.3.3). The third major conclusion of Experiment 2 was that model-based optimization could reduce
the error to levels comparable to the measurement precision for a stationary pupil. In other words, the foreshortening error induced by the rotation of the eye could be eliminated almost entirely. Such accurate correction is possible because the PFE surface is smooth and can be described with five parameters that can be estimated from calibration data, at least for artificial eyes.

On the basis of these empirical results we formulate three theoretical principles. They seem obvious in hindsight but were neglected in earlier PFE research (e.g., Brisson et al., 2013; Gagl et al., 2011). First, PFE analysis must be based on the three-way relationship between the eye, the fixation point on the screen, and the eye-tracking camera. Leaving the camera out of the equation leads to an ill-posed problem that does not permit a general solution. This is revealed when one compares data from different experimental layouts as was done here. Second, the foreshortening law is multiplicative in nature (Equation 5.3.1). Therefore, division rather than subtraction is the appropriate corrective operation (Equation 5.3.3), and the geometric rather than arithmetic mean is the appropriate measure of central tendency. Additive operations are justified only approximately for normalized data on the basis of the approximation \( \log(1 + x) \approx x \) for small \( x \). The third principle is the cosine relationship embedded in the geometric model (Equations 5.3.2 and 5.3.3) that grounds the PFE in the physics of the data acquisition process. With parameters that could have been fixed before any calibration data were collected, the geometric model accounted for over 82% of the variance in these data. In agreement with the first two principles, the model specifies a three-way multiplicative relationship between the coordinates of the eye, fixation target, and camera.

Our measurements of PFE are broadly consistent with published results that measured a single horizontal scan line using an artificial eye (Gagl et al., 2011), and a circular object pursuit task performed by human participants (Brisson et al.,
During the object pursuit task of Brisson et al. (2013), the pupil diameter systematically decreased as the object receded vertically and horizontally from the camera. The geometric model is consistent with this pattern of results. Gagl et al. (2011) reported a piecewise linear and quadratic function in their artificial eye measurements. Whereas our PFE maps exhibited a smooth quadratic trend, Gagl et al. (2011) reported that their artificial eye measurements showed a sharp linear increase near the left edge of the screen that then decreased quadratically as the gaze position moved rightward across the display. The quadratic pattern is consistent with the geometric model and predicts that the camera lens was on the left side of their EyeLink 1000 since the right eye position was measured. The drop in recorded pupil size near the left edge of the screen is anomalous and does not fit the geometric model predictions. One possible explanation is that there was a small imperfection in the artificial pupil that resulted in a loss of signal when the eye was oriented toward the left of the screen. We encountered a similar phenomenon with an early prototype of our artificial eyes. Specifically, there was a small spot inside the pupillary hole that the black paint missed and this spot resulted in a small hole in the thresholded pupil area in camera image for a few grid targets on the overall map. The resulting map was a smooth quadratic surface with an uncharacteristic drop in the region where this imperfection was visible to the camera. It seems plausible that the artificial eye of Gagl et al. (2011) had a similar imperfection that caused the apparent decrease in pupil size for fixations near the left margin of the screen. Alternatively, the different function obtained by Gagl et al. (2011) can stem from the fact that they used a tower-mounted EyeLink 1000, whereas we used a desktop configuration that placed the camera at a much greater distance from the eye.

One key objective of the current article is to provide the research community with a technique to correct for the PFE in a manner that does not sacrifice experimental
flexibility. The present methodology allows for unprecedentedly accurate PFE correction while preserving the freedom to study tasks such as reading or visual search that involve free viewing of the display. Both the parameter-free and the optimized geometric model reduced PFE enough that typical cognitive effects on the order of 5–10% should be detectable. One key advantage of the parameter-free geometric correction is that it does not require the researcher to collect any artificial-eye measurements. One weakness is that it does not account for the effects of the camera optics that alter the effective geometric layout. The optimized model virtually eliminates PFE by accounting for all sources of error, but requires calibration data to constrain the model parameters. For researchers that wish to use the optimized geometric model and use the EyeLink 1000, Table 5.3 lists the best-fitting parameters for the 3 layouts in our study. For researchers that use other eye trackers and/or layouts, the detailed description of our artificial eye and socket apparatus can help them calibrate their respective systems. The geometric relationships formalized in Equations 5.3.1–5.3.3 apply to all camera-based trackers, but the proprietary post-processing software of specific manufacturers may render these equations unsuitable for PFE correction purposes. Also, some of the simplifying assumptions used in the derivation of these equations will have to be re-examined for head-mounted trackers that put the camera much closer to the eye and thus the diameter of the eyeball is no longer negligible. These assumptions were made explicit in the Model section above.

The magnitude of the pupil foreshortening error depends greatly on the geometric configuration of the screen and camera relative to the eye (Figure 5.3). The following general recommendations can be used to inform the choice of layout. Putting the display far from the participant tends to reduce the range of oblique angles $\theta$ and hence the PFE. On the other hand, it increases the error in estimating the fixation location. This creates a trade-off between pupillometric accuracy and eye-tracking
accuracy because the latter deteriorates when the screen is far from the participant. An elegant solution is possible when the research question requires that the stimuli appear at multiple locations but does not constrain the spatial relationship among these locations. Our recommendation for such situations is to spread the stimuli along an arc that maintains a constant oblique angle $\theta$ with the camera-to-eye axis. The coordinates of these locations can be calculated from Equation 5.3.2 under the constraint $\cos \theta(x, y) = k = const$ for a suitably chosen $k$. The geometric model predicts identical PFE for all points along an arc of this type. Analogous circular contours of equal PFE are prominent in the calibration data (Figure 5.5). Another general recommendation is to place the camera as close to the eye-to-stimulus line of sight as the camera dimensions allow. Such placement makes $\cos \theta$ as close to 1 as possible. Desktop configurations provide more stable pupillometric data than tower-mounted and head-mounted configurations. Finally, we strongly discourage the participation of observers whose eyelids partially occlude their pupils. Glasses and contact lenses can also affect the PFE in unpredictable ways.

One potential limitation of our study is that we do not explicitly model the refracting effect of the cornea—the transparent curved layers in front of the iris (Atchison & Smith, 2000). The image recorded by video-based eye trackers measures the entrance pupil—the aperture stop of the iris as viewed through the refractive surfaces of the cornea. The refractive power of the cornea is considerable; it is approximately two times stronger than that of the lens of a relaxed eye (Atchison & Smith, 2000). The absolute magnification provided by the cornea does not affect the relative PFE. Rather, our concern is about corneal anisotropy. If different segments of the cornea have different refractive indices, the pupil diameter can appear to change when the eye rotates, turning different segments towards the camera. There are empirical data that bear on this issue (e.g., Spring & Stiles, 1948; Jay, 1962; Jennings & Charman,
1978). These studies photograph human eyes from a range of oblique angles, fit ellipses to the pupil images, and estimate the foreshortening coefficient as a function of the angle. A recent article (Mathur et al., 2013) re-analyzed the data from six published studies plus sophisticated new measurements. The overall conclusion was that, “Off-axis pupil shape is well described by a cosine function that is both decentered by a few degrees [towards the temporal side] and flatter by about 12% than the cosine of the viewing angle.” (Mathur et al., 2013, p. 7).

The empirical foreshortening function of Mathur et al. (2013) can easily be incorporated into our geometric model. To that end, we substitute their Equation 9 for the term $\cos \theta$ in our Equation 5.3.3, yielding:

$$\phi_0 = \frac{\phi(x, y)}{\sqrt{0.992 \cos \left( \frac{\theta(x,y)+5.3}{1.121} \right)}} \quad (5.4.1)$$

where $\theta(x, y)$ is determined from Equation 5.3.2, in degrees. Note that the data of Mathur et al. (2013), as well as the earlier data re-analyzed therein (Spring & Stiles, 1948; Jay, 1962; Jennings & Charman, 1978, etc.), were all collected along the horizontal equatorial line. One area for future research would be to systematically map the apparent pupil foreshortening across the entire visual field in participants with dilated eyes, and compare the human eye data to our artificial eye data to quantify the effect of corneal refraction.

Pupil foreshortening error is a potentially large confound that must be taken into account prior to interpreting pupillary data from image-based eye trackers. We introduced an artificial eye and socket model and systematically mapped the PFE across the horizontal and vertical visual field in three separate experimental layouts. In light of previous cognitive pupillometric research, our data strongly indicate that the PFE is larger than many cognitive pupillometric effects. The PFE is not currently corrected by popular commercial image-based eye tracking systems. We formulated
a simple parameter-free geometrical model that reduced the root mean squared error of the PFE by 82.5% and an optimized model that reduced it by 97.5%. Thus, very accurate PFE correction is possible and the corrected pupillometric data have the precision necessary to measure typical cognitive effects without limiting the types of tasks that can be studied by researchers.
<table>
<thead>
<tr>
<th>Layout</th>
<th>RMSE</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggregate data</td>
<td>0.034</td>
<td>0.913</td>
<td>1.056</td>
</tr>
<tr>
<td>Model multipliers (PF)</td>
<td>0.029</td>
<td>0.925</td>
<td>1.046</td>
</tr>
<tr>
<td>Corrected data (PF)</td>
<td>0.006</td>
<td>0.987</td>
<td>1.014</td>
</tr>
<tr>
<td>Model multipliers (opt.)</td>
<td>0.034</td>
<td>0.912</td>
<td>1.056</td>
</tr>
<tr>
<td>Corrected data (opt.)</td>
<td>&lt;0.001</td>
<td>0.998</td>
<td>1.002</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Layout</th>
<th>RMSE</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggregate data</td>
<td>0.025</td>
<td>0.941</td>
<td>1.040</td>
</tr>
<tr>
<td>Model multipliers (PF)</td>
<td>0.025</td>
<td>0.944</td>
<td>1.038</td>
</tr>
<tr>
<td>Corrected data (PF)</td>
<td>0.004</td>
<td>0.993</td>
<td>1.009</td>
</tr>
<tr>
<td>Model multipliers (opt.)</td>
<td>0.025</td>
<td>0.941</td>
<td>1.039</td>
</tr>
<tr>
<td>Corrected data (opt.)</td>
<td>&lt;0.001</td>
<td>0.998</td>
<td>1.002</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Layout</th>
<th>RMSE</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggregate data</td>
<td>0.020</td>
<td>0.949</td>
<td>1.033</td>
</tr>
<tr>
<td>Model multipliers (PF)</td>
<td>0.019</td>
<td>0.952</td>
<td>1.030</td>
</tr>
<tr>
<td>Corrected data (PF)</td>
<td>0.004</td>
<td>0.992</td>
<td>1.008</td>
</tr>
<tr>
<td>Model multipliers (opt.)</td>
<td>0.020</td>
<td>0.947</td>
<td>1.033</td>
</tr>
<tr>
<td>Corrected data (opt.)</td>
<td>&lt;0.001</td>
<td>0.998</td>
<td>1.002</td>
</tr>
</tbody>
</table>

Table 5.2 Descriptive statistics of the deviation in pupil diameter from the geometric mean for the aggregate layout maps, the corrective multipliers generated by the parameter-free (PF) model and the model with optimized (opt.) parameters, and the corrected data (parameter-free and optimized). RMSE = root mean squared error.
<table>
<thead>
<tr>
<th>Model Parameters</th>
<th>Near</th>
<th>Medium</th>
<th>Far</th>
</tr>
</thead>
<tbody>
<tr>
<td>Camera-to-pupil, $C_x$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical</td>
<td>92</td>
<td>92</td>
<td>92</td>
</tr>
<tr>
<td>Optimized</td>
<td>130</td>
<td>165</td>
<td>183</td>
</tr>
<tr>
<td>Camera-to-pupil, $C_y$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical</td>
<td>-310</td>
<td>-310</td>
<td>-310</td>
</tr>
<tr>
<td>Optimized</td>
<td>-215</td>
<td>-239</td>
<td>-230</td>
</tr>
<tr>
<td>Camera-to-pupil, $C_z$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical</td>
<td>495</td>
<td>525</td>
<td>625</td>
</tr>
<tr>
<td>Optimized</td>
<td>$495^*$</td>
<td>$525^*$</td>
<td>$625^*$</td>
</tr>
<tr>
<td>Screen-NW-corner, $S_x$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical</td>
<td>-163</td>
<td>-163</td>
<td>-163</td>
</tr>
<tr>
<td>Optimized</td>
<td>-142</td>
<td>-87</td>
<td>-76</td>
</tr>
<tr>
<td>Screen-NW-corner, $S_y$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical</td>
<td>58</td>
<td>58</td>
<td>58</td>
</tr>
<tr>
<td>Optimized</td>
<td>206</td>
<td>140</td>
<td>156</td>
</tr>
<tr>
<td>Screen-NW-corner, $S_z$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical</td>
<td>740</td>
<td>835</td>
<td>935</td>
</tr>
<tr>
<td>Optimized</td>
<td>736</td>
<td>851</td>
<td>937</td>
</tr>
</tbody>
</table>

Table 5.3 Comparison of geometric models (parameter-free physical measurements vs. optimized parameters) for Near, Medium, and Far Layouts. All values are distances in mm. $C_x$, $C_y$, $C_z$ are the coordinates of the camera; $S_x$, $S_y$, $S_z$ are the coordinates of the upper left-hand corner of the screen (cf. Figure 5.3). An asterisk indicates the parameter was fixed to the physical layout measurement.
CHAPTER 6

Relational Foraging: Tracking exploration-exploitation trade-offs during relational reasoning

6.1 Introduction

The ability to adaptively regulate the balance between exploration and exploitation is critical for optimizing behavior in the diverse, dynamic environments we encounter on a daily basis. Whether it is deciding to have lunch at your favorite restaurant in lieu of trying a new restaurant or discontinuing your current line of research in favor of exploring new ideas, the decision to exploit known sources of reward or explore new options is faced for decisions big and small across virtually all time-scales. Despite the ubiquity of the exploration-exploitation tradeoff and its broad importance in decision making and executive control, the neural mechanisms involved are still not well understood (Cohen, McClure, & Yu, 2007). Promising work combining animal recordings (Aston-Jones et al., 1994; Hollerman & Schultz, 1998; Hasselmo & Cekic, 1996) and computational modeling (Usher et al., 1999; Brown, Gilzenrat, & Cohen, 2005; Montague, Hyman, & Cohen, 2004; Yu & Dayan, 2005) have begun to outline a specific set of neural systems and mechanisms that may be involved—of these systems the locus coeruleus (LC) noradrenergic system is thought to play a central role (Aston-Jones & Cohen, 2005; Cohen et al., 2007). Despite the rich
literature on LC function in animals, there have been limited human studies testing the theorized link between LC function and the exploration-exploitation trade-off in humans (Gilzenrat et al., 2010; Jepma & Nieuwenhuis, 2011). Here we combine pupillometry as an indirect index of LC activity and concurrent verbal protocols to tag exploratory and exploitative solution periods to provide the first evidence that the LC mediates real-time shifts in exploration and exploitation during analogical reasoning.

The current working framework underlying the exploration-exploitation trade-off is thought to rely on the interplay between several neuromodulatory systems including the midbrain dopaminergic system (Montague et al., 2004), the basal forebrain cholinergic system (Yu & Dayan, 2005), and the LC noradrenergic system (Aston-Jones & Cohen, 2005; Yu & Dayan, 2005). The dopamine system is thought to play an important role in learning the mapping between agent actions and rewards within the environment, while the cholinergic system monitors expected uncertainty within the environment. The LC noradrenergic system is thought to monitor for unexpected uncertainty and actively mediate the shift between exploration and exploitation in response to reward history (Aston-Jones & Cohen, 2005). Much of the current theory of LC function is based on electrophysiological animal recordings (See Aston-Jones & Cohen, 2005 for review) and computational models of this empirical data (Usher et al., 1999; Brown et al., 2005). In particular, the electrophysiological data revealed that LC neurons in monkeys exhibit distinct patterns of neuronal firing that lie along a continuum of task performance: offline, tonic, and phasic mode broadly construed. In offline mode a low level of LC base-rate firing is observed and occurs if the animal is drowsy or sedated, producing low task performance. In tonic mode a high level of LC base-rate firing is observed and the animal exhibits an exploratory state of engagement in which they are sensitive to task-related and task-unrelated stimuli,
resulting in low task performance. In phasic mode, base-rate LC firing is low and LC neurons fire bursts of activity synced to task-relevant events and the animals exhibit high task performance. These findings have been incorporated into a broader theory of LC function (Adaptive Gain Theory AGT; Aston-Jones & Cohen, 2005). AGT posits that low tonic baseline LC activity promotes exploitation and high tonic baseline LC activity promotes exploration. However, this component of AGT was extrapolated as the simple tasks the monkeys performed in the electrophysiological studies did not strongly promote exploration-exploitation trade-offs and there was a lack of supporting human studies.

A major obstacle to studying LC function in humans is identifying a non-invasive method for measuring LC activity. Recently, pupil diameter has emerged as a promising noninvasive proxy measure for LC activity, and is being increasingly employed for this purpose (e.g., Cheadle et al., 2014; Einhäuser et al., 2010; Wolfgang et al., 2008; Eldar et al., 2013; Gilzenrat et al., 2010; Jepma & Nieuwenhuis, 2011). Neuroimaging work has shown that fMRI BOLD LC activity and pupillary response covary (Murphy et al., 2014) and that both the P3 event-related potential and pupil diameter track are sensitive to LC-NE modes of task engagement (Murphy et al., 2011; Cheadle et al., 2014). Additional evidence from electrophysiology (e.g., Rajkowski et al., 1994) and pharmacology (e.g., Koss, 1986; Phillips et al., 2000) suggests that pupil diameter also correlates with LC activity in animals. The anatomical pathways linking the locus coeruleus and the pupil are a topic of ongoing research, but probably involve α2-adrenoreceptor mediated inhibition of the parasympathetic Edinger-Westphal nucleus responsible for pupil constriction (Samuels & Szabadi, 2008a, 2008b).

Here we extend the study of the exploration-exploitation trade-off and LC function by tracking real-time shifts in exploration and exploitation in a rich, temporally extended analogical reasoning task, Raven’s Advanced Progressive Matrices (APM;
Raven et al., 1998). Raven’s APM is an excellent environment to induce shifts between exploration and exploitation as it actively manipulates difficulty and environment familiarity by repeatedly placing participants in an unfamiliar relational environment in which they must engage in “relational foraging” to attempt to arrive at the correct answer. The pupillary response was used as an indirect proxy for LC activity and was collected for each trial. In order to allow for the identification of exploratory and exploitative periods concurrent think aloud verbal protocols were recorded as participants solved each item. The novel combination of pupillometry and verbal protocol analysis revealed a decrease in pupillary response during exploitative periods and boost during exploratory periods that increased with fluid ability—providing the first empirical evidence, albeit indirect, that the LC supports real-time shifts in the exploration-exploitation tradeoff during analogical reasoning in humans.

6.2 Method

Twenty university students with normal or corrected-to-normal vision completed a short-form test from Raven’s Advanced Progressive Matrices, Set II (Raven et al., 1998). Participants either completed items 2, 4, 6, 9, 10, 11, 16, 17, 19, 21, 23, 24, 26, and 29 or items 1, 3, 5, 7, 12, 13, 14, 15, 18, 20, 22, 25, 27, and 28. The two short-form test sets were matched by item difficulty level and relation type. The instructions followed the Raven APM Manual guidelines for individual test administration (Raven et al., 1998). Concurrent “think aloud” verbal protocols were collected according to standard procedures using adapted think aloud instructions from (Ericsson & Simon, 1993). After participants received instructions on thinking aloud, they practiced on unrelated items like thinking aloud while they completed multiplication problems.
Figure 6.1  *Raven problem format and trial sequence.* Left: The problem matrix and the 8 response alternatives are shown with solid lines. The height of the rectangular box around the matrix subtended 9 degrees of visual angle. This example item (generated by the authors) contains three relations that must be extracted: distribution of three shapes (diamond, triangle, parallelogram), distribution of three line orientations (0°, 45°, 90°), and decreasing line number down columns (3→2→1). Right: Each trial had three phases: fixation, solution, and response. Participants fixated for 1 s. Eye movements and concurrent think aloud verbal protocols were collected during the solution phase. Moving the mouse cursor out of the fixation box triggered the response phase, during which the problem matrix was masked and the participant clicked on their chosen answer. The inter-trial interval (ITI) was 200 ms.

until the experimenter was confident they understood the task. Participants were paid $6 per hour plus $1 bonus for each correct answer.

Stimuli were presented at 60 Hz on a 21” NEC AccuSync 120 color CRT and were viewed binocularly from a head rest located 92 cm away. Each trial began with a brief alert sound and a fixation cross appeared in the middle of the screen (Figure 6.1, right). After the participant fixated for 1 s, which allowed for equipment re-calibration, the Raven problem appeared and the participant had unlimited time to work on it. A mouse click on one of the responses ended the trial. No accuracy
feedback was provided during the course of the study to control for feedback-induced pupillary responses. Accuracy and solution time were collected for each trial. Accuracy was defined as the total number of items answered correctly and solution time was measured from stimulus onset until response selection.

Eye-tracking data protocols were also collected for each APM item. The pupil area of the left eye was measured using an Eyelink 1000 desktop eye tracker (SR Research, 2006) at a sampling rate of 250 Hz. The experimental room had a consistent ambient illuminance with 25 lux incident at participants’ eyes to control for the pupillary light reflex. Image analysis of the Raven APM items revealed high luminance consistency across test items and across the problem and response cells within each item. Therefore, we were able to avoid altering the properties of the original Raven APM test images and maintain their original psychometric properties.

6.2.1 Pupil Data Preprocessing and Gaze Position Correction

Pupillary data was corrected for blink artifacts and then corrected for gaze position pupil error. Pupil diameter was recorded from pre-trial fixation through the end of the inter-trial-interval. Following standard procedures pupillary measurements were filtered for blink artifacts, linearly interpolated, and then smoothed for measurement noise (Beatty & Lucero-Wagoner, 2000; Klingner, 2010). In addition to these standard procedures our pupil data had to be corrected to account for changes in gaze position that result in systematic measurement error due to foreshortening of the pupillary image as the eye rotates away from the camera (Hayes & Petrov, 2014a). The pupil foreshortening error correction described in detail in Hayes and Petrov (2014a) fits a geometric model that estimates the foreshortening effect as a function of the cosine of the angle between the eye-to-camera axis and the eye-to-stimulus
axis. The geometric correction reduced the root mean squared error in pupil diameter from an artificial eye with fixed pupil diameter by 97.5%. The results strongly indicated that the pupil foreshortening error is invariant across changes in pupil size and systematically varies as a function of the orientation of the eye with respect to the camera (Hayes & Petrov, 2014a). In addition their results corresponded well with previous empirical measurements of pupil foreshortening error in biological human eyes (Spring & Stiles, 1948; Jay, 1962; Jennings & Charman, 1978; Mathur et al., 2013). Together these findings suggest the geometric correction procedure can be employed to minimize pupil foreshortening error in our current data due to changes in gaze position during relational reasoning. The geometric model correction from Hayes and Petrov (2014a) was applied to correct for pupil foreshortening error following artifact correction.

### 6.2.2 Verbal Protocol Coding Procedure

The concurrent think aloud verbal protocols were used to segment each trial into exploratory and exploitative solution periods. A broad coding scheme was developed to assist the coder in identifying exploratory and exploitative content during Raven solution. A semi-automated coding routine was developed in Matlab that was used to code all verbal protocol data. In this routine, for each trial the coder would be presented with an image of the current APM item while the recorded verbal protocol audio was played back in real-time. The coder served as an ‘exploration detector’ pressing the J key to indicate the beginning of an exploratory period and the K key to indicate the end of an exploratory period and the beginning of an exploitative period. The beginning of a trial was coded as neutral prior to any key presses. The semi-automated coding scheme would then convert the key press time stamps into a verbal protocol code stream that contained the neutral (0), exploratory (+1) and
Figure 6.2 *Mean change in pupil diameter by fluid intelligence and*  
Left: Comparisons of the mean percent change in pupil diameter from baseline for exploratory and exploitative periods is shown averaging across all participants and for subgroups of participants with APM scores of less than 8, 8-10, 11-12, and 13-14 (N=5 for each subgroup). The error bars represent the standard error of the mean. Right: Scatterplot of APM score and mean % change in pupil diameter during exploration. A linear regression revealed that mean exploratory pupil diameter explained 25% of the variance in APM score.

exploitative (-1) codes for that trial, each sampled at 250 Hz. This procedure was completed for all participants (N=20) and trials (N=14) resulting in 280 individual code streams.

### 6.2.3 Pupil and Verbal Protocol Analysis

In order to synchronize the pupillary response stream with the verbal protocol code stream three sources of latency were considered: participant latency, coder latency, and LC/pupillary response latency. Participant latency refers to the latency that occurs due to a participant processing the APM item information and transforming
it into an utterance. Participant latency unfortunately cannot be accounted for in our study as it is known to vary across individuals and types of processing steps and, therefore, will invariably add some noise to our data (Ericsson & Simon, 1993). Coder latency and LC pupillary response can and were accounted for prior to analysis. Coder response latency refers to the processing time it takes for the coder to process what they are hearing, make the decision to switch codes, and then actually press the key on the keyboard. To estimate this value, a random sample of 50 trials was used to compare the coder key response time stamps to the original audio time series using audio editing software (Apple, 2010). The results showed a coder response latency of approximately 1 second (M=1014 ms, SD=198 ms). Finally, we considered the documented lag between LC activity and the pupillary response. Single cell studies of LC neurons show LC activity is tightly linked to stimulus onset lagging approximately 200 ms (Rajkowski, Majczynski, Clayton, & Aston-Jones, 2004; Clayton, Rajkowski, Cohen, & Aston-Jones, 2004). However, the temporal resolution of the pupillary response is much lower than LC neurons, acting as a low-pass filter of LC activity with a lag of approximately 1 second after stimulus onset (Hayes & Petrov, 2014b; Gagl et al., 2011; van Steenbergen & Band, 2013). Since the coder and pupillary response latencies were approximately equivalent (each about 1 second) no additional preprocessing was necessary to synchronize the pupil and code streams prior to analysis.

The concurrent think aloud protocols were used to segment each trial into exploratory and exploitative periods based on their content. Pupillary response was computed for each exploratory or exploitative segment as the percent change in pupil dilation (task-evoked pupil diameter minus baseline pupil diameter, divided by baseline diameter) for each subject and each trial. The baseline for the first segment at the beginning of each trial was computed using the first 500 ms of that segment to
provide a more accurate baseline estimate as participants began the trial. The baseline for all subsequent trial segments was computed as the average pupil diameter during the 1000 ms in the preceding segment.

6.3 Results

Mean error rate and mean solution time both increased as a function of trial in agreement with previous findings (Bors & Vigneau, 2003). A one-tailed Pearson’s product-moment correlation test revealed the solution time trial effect was slightly larger, but trial accounted for a significant amount of variance in both measures (error rate, \( t(12)=6.55, p<.0001, r^2=.78 \); solution time, \( t(12)=8.46, p<.0001, r^2=.86 \)).

The error rate and solution time effects are not surprising as APM item difficulty increases as trial number increases. In order to examine whether solution time varied as a function of fluid ability, participants were split into four subgroups (\( N=5 \)) based on their APM score: high ability (\( HGf \), scores 13-14), medium-high ability (\( MHGf \), scores 11-12), medium-low ability (\( MLGf \), scores 8-10) and low ability (\( LGf \), scores < 8). A split-plot ANOVA was performed to determine whether solution time varied significantly across fluid ability groups, with group as a between-subjects factor and trial as a within-subjects. The results reconfirmed a significant main effect of trial on solution time, \( F(13, 208)=11.32, p<.0001, \eta_p^2=.41 \), but more importantly indicated that there was not a significant main effect of group, \( F(3, 16)=1.87, p=.18 \), or trial by group interaction, \( F(39, 208)=1.38, p=.08 \). This suggests that higher fluid ability participants did not spend significantly more time solving the APM items than lower ability participants.

The overall number of exploration-exploitation switches increased as APM item difficulty increased. A trend analysis across all participants confirmed a significant linear trend, \( F(1, 247)=108.99, p<.0001, \eta_p^2=.31 \), and quadratic trend, \( F(1, 247)=12.04, \eta_p^2=.08 \).
\( p < .001, \eta_p^2 = .05, \) in the total number of switches as a function of trial. On early trials participants often only exhibited one exploration-exploitation switch, but on the more difficult middle and late APM items participants would often shift multiple times between exploring and exploiting. A split-plot ANOVA with group as a between-subjects factor and trial as a within-subjects factor showed no main effect of group, \( F(3, 16) = 0.30, p = .82, \) or trial by group interaction, \( F(39, 208) = 1.16, p = .25, \) on transition number. A similar pattern was observed in the difference in time duration spent exploring versus exploiting. A split-plot ANOVA revealed a significant increase in time exploring as trial difficulty increased, \( F(13, 208) = 2.28, p < .01, \eta_p^2 = .12, \) but no group effect, \( F(3, 16) = 1.53, p = .24, \) or trial by group interaction, \( F(39, 208) = 0.78, p = .82. \) To summarize, while the overall number of exploration-exploitation switches and time spent exploring versus exploiting increased with trial difficulty, no significant fluid ability or fluid ability by trial difficulty interaction effects were observed in either dependent measure. These findings in conjunction with the solution time findings suggest that the observed differences in fluid ability are not well explained by solution time, the number of exploration-exploitation switches, or the difference in the overall amount of time spent exploring or exploiting.

A significant boost in mean pupillary response was observed during exploratory solution periods relative to exploitative solution periods (Figure 6.2, left, one-tailed paired, \( F(1, 265) = 111.51, p < .0001, r^2 = .29). \) This broad finding provides novel indirect support that the LC-NE system mediates the exploration-exploitation tradeoff during analogical reasoning. In order to examine whether exploratory boost in pupil diameter across groups was significant a trend analysis was performed. The trend analysis revealed that mean exploratory pupillary response increased linearly as a function of group \( (F(1, 15) = 5.57, p = .03, \eta_p^2 = .27). \) Using mean exploratory pupil diameter as a predictor variable in a linear regression explained 25% of variance in
individual differences in APM score (Figure 6.2, right). This is another novel finding suggesting that the mediation of the exploration-exploitation tradeoff may contribute to individual differences in fluid intelligence. No significant trends were observed in exploitative pupillary response as a function of fluid ability. Finally, we performed a trend analysis across all participants to determine whether exploratory or exploitative pupillary response were modulated by trial difficulty. The trend analysis revealed a linear decrease in mean pupillary response during exploratory periods as a function of trial difficulty ($F(1, 234)=27.98, p<.00001, \eta_p^2=.11$) and no change during exploitative solution periods ($F(1, 234)=1.06, p=.30$). This decrease in exploratory pupillary response across trials is likely related to the increase in overall time spent exploring across trials as discussed in the previous section.

6.4 Discussion

Previous research using the pupillary response to index the LC-NE system to study the exploration-exploitation tradeoff has been limited by experimental designs that did not strongly promote exploration (Gilzenrat et al., 2010) or had potential confounding factors (Jepma & Nieuwenhuis, 2011). Gilzenrat et al., 2010 had participants complete an auditory pitch discrimination task in which reward increased as the pitch discrimination difficulty increased, until the discrimination eventually became impossible. Importantly, participants were allowed the option to escape prior to each trial, which would reset the reference tone, difficulty, and reward levels. They found that baseline pupil diameter increased leading up to escape trials and decreased afterwards consistent with our findings of a decrease during exploitation and increase during exploration. However, the effect size was modest and the authors suggested this may be due to the escape manipulation not sufficiently emulating exploration. In
a followup study, Jepma & Nieuwenhuis, 2011 tracked exploration-exploitation trade-offs during a dynamic n-armed bandit gambling task in which participants repeatedly had to choose to play one of four slot machines with non-stationary rewards. While the dynamic n-armed bandit task more strongly promoted shifts in the exploration-exploitation tradeoff, gaze position was not controlled during bandit selection and reward feedback was visually presented immediately following selection, restricting their analysis to the pre-trial baseline period. The main result showed an overall increase in pupil baseline prior to exploratory trial choices (i.e. trials where participants switched their slot machine choice) compared to exploitative trial choices (i.e. trials in which participants picked the same slot machine). Our study supplements these earlier studies by strongly promoting shifts in the exploration-exploitation tradeoff and limiting confounding factors (i.e. feedback and pupil foreshortening error). Taken together these three studies provide converging indirect support for the role of the involvement of the LC-NE system in the exploration-exploitation tradeoff.

The present work also expands the domain in a novel direction by examining how individual differences in fluid intelligence covary with pupillary response during control state shifts. Recent work has indicated that high fluid intelligence individuals have larger task-evoked pupillary responses when performing difficult tasks lending support to the view that high Gf individuals may simply have more cognitive resources that can be recruited for more demanding tasks (Van Der Meer et al., 2010). Earlier work showed the opposite pattern in which higher intelligence individuals showed smaller task-evoked pupillary responses than average intelligence individuals, lending support to the hypothesis that high intelligence individuals use their cognitive resources more efficiently (Ahern & Beatty, 1979, 1981). Our results do not directly refute either of these earlier views, but seem to offer a third account. Specifically, that higher fluid ability individuals are able to more effectively regulate their task-relevant
control state and how their available cognitive resources are deployed. Exemplified by our finding that the exploratory boost in pupil diameter was modulated by fluid intelligence, this opens up the interesting possibility that individual differences in fluid ability may be related to individual differences in mediating control state through stronger shifts in neural gain. In addition the deployment hypothesis offers a possible explanation for the conflicting earlier findings on the relationship between intelligence and pupillary response. In tasks that require exploration (such as the visual analogy task used by Van Der Meer et al. (2010)) high Gf individuals that shift into higher-gain states will have larger task-evoked pupillary responses than lower Gf individuals. Moreover, on overlearned tasks that primarily require exploitation (such as the mental multiplication, digit span used by Ahern and Beatty (1979, 1981) ) are easier for higher Gf individuals than lower Gf individuals resulting in lower pupillary response for high Gf individuals. While our study does not directly bear on pupillary response during overlearned tasks, there are many studies indicating easier tasks induce smaller pupillary response than difficult tasks (See Beatty & Lucero-Wagoner, 2000 for a review of this work).

In conclusion, our unique combination of pupillometry and verbal protocol analysis was used to identify and track shifts in the exploration-exploitation tradeoff while participants completed items from the APM fluid intelligence test. The results showed decreased pupillary response during exploitation and increased pupillary response during exploration consistent with prominent theories of LC-NE function. Importantly, individual differences in fluid ability were linearly related to the increase in pupillary response during exploratory periods. To our knowledge this is the first study showing evidence that noradreneric activity may mediate the exploration-exploitation tradeoff during analogical reasoning and the only study to show an exploration-related boost in pupillary response that is modulated by individual differences in fluid ability.
CHAPTER 7

General discussion

Many cognitive scientists have argued that analogical ability is a foundational aspect of human cognition (French, 2002; Hofstadter, 1995). Therefore, any research that sheds even a small amount of light on relational reasoning has the potential to provide a myriad of advances into how we understand the human cognitive system as well as the development of technologies that can harness the power that relational ability affords. Using the Raven Advanced Progressive Matrices test (APM) as a prototypical testing ground, the present studies combined traditional measures such as total APM score, solution time, and eye movements with two new techniques (Successor Representation Scanpath Analysis, SRSA; Pupil Foreshortening Error correction, PFE) to understand important mechanisms of visual relational reasoning (i.e. attentional control, attentional scope, and the exploration-exploitation tradeoff) and how they contribute to individual differences in fluid intelligence.

Chapter 2 provided new insights into individual differences in relational reasoning ability by overcoming a major methodological obstacle in understanding attentional control during APM solution. Existing methods for studying attentional control using eye movements were not able to reliably predict Raven performance because they either struggled with the extended scanpaths generated by Raven performance (i.e. string edit and transition probability methods) or were poorly suited for exploratory data analysis (i.e. Hidden Markov models). By pioneering the use of reinforcement
learning to analyze eye movements via SRSA, interpretable processing strategies were extracted and predicted 41% of the variance in individual differences in APM score. Row-wise scanning was associated with higher APM scores, while toggling between problem and response information was associated with lower APM scores. The SRSA toggle component replicated earlier findings of negative correlations with toggling and APM performance (Carpenter et al., 1990; Bethell-Fox et al., 1984; Vigneau et al., 2006), but was more sensitive than traditional measures (e.g. number of toggles). The systematicity was a novel finding providing the most detailed picture of APM performance and the importance of strategic attentional processing to date. More generally, the novel SRSA technique permits access and greater sensitivity to sequential eye movement information that could prove useful in a wide array of other research and practical applications.

Applying SRSA to the area of cognitive training in Chapter 3 resulted in a number of advancements. While previous studies have regularly observed test-retest practice effects on the Raven APM, they have failed to offer any account of what is actually being refined with practice (Bors & Vigneau, 2003). Using the difference in test-retest SRs to predict score-gains on the Raven APM predicted approximately a third of the variance in score-gains and indicated refinements in strategy (i.e. increases in row-systematicity and careful answer checking) is a major reason why scores on the APM improve with practice. This finding advances our understanding of the Raven APM test and individual differences in performance in test-retest designs using the Raven APM. Perhaps more importantly, it raises the possibility that strategy refinement may be a serious confound in the gF enhancement literature. Specifically, strategy refinement when considered in light of the positive findings showing control-adjusted gains in gF for visual training regimens, but not for non-visual training regimens (Stephenson & Halpern, 2013; Jaeggi et al., 2014), suggests that the visual training
regimens may be inadvertently training information processing strategies that can then be applied to Raven at posttest. Therefore, our findings raise significant doubts about common practices in the rapidly growing field of gF enhancement and argue for the use of a wider battery of both visual and non-visual training regimens and gF measures, as well as the inclusion of SRSA to monitor and control for strategy refinement.

In Chapter 4 the presentation of relational information was actively manipulated and provided new insights into the roles of processing strategy and working memory during relational reasoning. In Experiment 1 the presentation order of relational information was actively manipulated to be either purely systematic or haphazard. The results revealed novel support to the hypothesized link between information processing strategy and the probability of relational insight (Grant & Spivey, 2003; Thomas & Lleras, 2007). Specifically, systematic scanning produced a boost in relational performance and haphazard scanning produced a significant deficit in relational reasoning ability. Experiment 2 examined the effects of reduction in problem information available to VSTM and iconic memory during relational reasoning using a gaze-contingent viewing paradigm in which peripheral vision was eliminated. A large reduction in information available to VSTM and iconic memory did not affect relational ability. These results suggest that unattended peripheral information plays only a minor role during relational reasoning and suggests that participants deploy their attentional resources on a cell-by-cell basis. The results of Experiment 2 in addition to advancing our understanding of the role working memory systems during relational reasoning, also have applied value because the gaze-contingent viewing condition simulates tunnel vision caused by many forms of degenerative eye disease (e.g. glaucoma, retinitis pigmentosa, cataracts). This is of importance because these degenerative eye diseases affect hundreds of millions of people worldwide (Pascolini
& Mariotti, 2012). Our results suggest that complex relational reasoning can still be performed within these clinical populations.

Chapter 5 advanced our understanding of pupil foreshortening error (PFE) and demonstrated how it can be virtually eliminated using a simple geometric model. Experimental designs for controlling PFE and preexisting methods for correcting PFE would not allow for cognitive pupillometry in free-viewing tasks like Raven, therefore a better correction procedure was needed. Experiment 1 empirically demonstrated a strong ratio scale between the arbitrary units reported by the EyeLink 1000 eye tracker and physical units. This is an important advancement as a ratio scale is necessary in order for multiplicative relationships to be meaningful and to our knowledge, this was not previously empirically established. Experiment 2 systematically mapped PFE across the full display for three different experimental layouts using three artificial eye models with different pupil sizes. The empirical results indicated large PFE across all 9 empirical maps and strongly suggested that pupillary effects should not be cognitively interpreted if PFE is not corrected. The measured PFE was corrected by a geometric model that expressed the foreshortening of the pupil area as a function of the cosine of the angle between the eye-to-camera axis and the eye-to-stimulus axis. The model reduced the root mean squared error of pupil measurements by 82.5% when the model parameters were pre-set to the physical layout dimensions, and by 97.5% when they were optimized to fit the empirical error surface. This critical advancement made it possible to use pupillary response to study the exploration-exploitation tradeoff during relational reasoning in Chapter 6. More broadly, the geometric model correction opens the door for using cognitive pupillometry in a wide range of other important areas that require changes in gaze position (e.g. reading, visual search).
Finally, Chapter 6 used a novel combination of verbal protocol analysis and pupillometry to advance our understanding of the role of exploration-exploitation tradeoffs during relational reasoning. Previous research using the pupillary response to index the LC-NE system during exploration-exploitation tradeoffs has been limited by experimental designs that did not strongly promote exploration (Gilzenrat et al., 2010) or had potential confounding factors (Jepma & Nieuwenhuis, 2011). By pioneering the use of pupillometry to study the exploration-exploitation tradeoff during relational reasoning, we were able to provide the first evidence that the pupillary response tracks real-time shifts in the exploration-exploitation tradeoff during analogical reasoning. The increase in pupil diameter during exploration and the decrease in pupil diameter during exploitation lends novel support to prominent theories that suggest the LC-NE system is important for mediating the exploration-exploitation tradeoff. The work is also the first to show evidence that pupillary response during exploration explains 25% of the individual differences in APM score. This finding opens up the interesting possibility that individual differences in fluid intelligence may be related to individual differences in mediating control state through stronger shifts in neural gain during exploration. It also offers a possible explanation for earlier contradictory findings that higher intelligence individuals have smaller task-evoked pupillary responses during mental multiplication and digit span tasks (efficiency hypothesis, Ahern & Beatty, 1979, 1981) and higher intelligence individuals have larger task-evoked pupillary responses during visual analogy tasks (resource hypothesis, Van Der Meer et al., 2010).

The results of the present work have advanced our understanding of the mechanisms of visual relational reasoning and individual differences in fluid intelligence in a number of ways. The development of SRSA allowed access to detailed and interpretable information from sequential eye movement data on the strategic allocation
of attentional control during relational reasoning. This work revealed a surprisingly strong role for strategic processing that explained individual differences in relational reasoning on Raven with unprecedented precision. SRSA also revealed that strategy refinement explained one third of the variance in test-retest practice effects on Raven—revealing strategy refinement as a potentially serious confound within the cognitive training literature. Active manipulations of strategic processing and the amount of information available to working memory systems served to directly quantify the relationships between strategy, working memory, and relational reasoning. Finally, through the development of a new method for correcting pupil foreshortening error we were able to assess real-time shifts in the exploration-exploitation tradeoff during relational reasoning by combining pupillometry and verbal protocol analysis. The findings revealed important new insights into the contribution of the exploration-exploitation tradeoff to relational ability, pupillary response as non-invasive index of LC-NE activity, and the relationship between exploratory noradrenergic activity and individual differences in fluid intelligence.

7.1 Future Directions

The studies detailed here have provided a number of advancements in our understanding of individual differences in relational reasoning, but they also raise new questions and open up avenues for future work. Arguably the research area that could have the broadest potential impact is providing direct evidence that strategy refinement explains Gf gains in pretest-train-posttest designs with visual training regimens. The field of cognitive enhancement is a rapidly growing area in which a large amount of resources are being deployed. Much of the focus to date has been on the development of different types of training regimens while the assessment of change in cognitive ability has arguably lagged far behind. Chapter 3 took the first
step in understanding test-retest effects and revealed indirect evidence for strategy refinement as a serious confound within the cognitive training literature. The next step will be to provide direct evidence on whether strategy refinement can explain control-adjusted gains when visual training regimens are used. The proposed study will consist of a pretest-train-posttest design with visual N-back WM training and an active control group. SRSA will be used to remove the effects of strategy refinement and determine if any remaining gain in Gf is observed. If a significant training effect disappears after strategy refinement is removed, this will be an important study that will help explain the mixed cognitive enhancement results of the past and move the field toward more thoughtful ways of assessing cognitive gains in the future. On the other hand, if a significant training effect remains this will provide compelling new evidence arguing in favor of the efficacy of cognitive training and will help justify the large amount of resources being invested.

The targeted gaze-contingent manipulation in Chapter 4 revealed no effects of constraining attention during relational reasoning across all participants. However, individual differences may still exist in how attentional resources are deployed that are masked by this aggregate analysis. An area for future research is to analyze individual differences in visual attention during relational reasoning. A study which combines latent variable measures of attentional capacity and attentional control with a within-subjects normal vs. gaze-contingent manipulation of Raven could prove successful in separating out the unique contributions of attentional capacity and dynamic updating of attentional scope to individual differences in fluid intelligence. An additional study which systematically changes the size of the gaze-contingent window could prove useful in identifying potential differences in the spatial extent of attention in high and low Gf participants.
Chapter 6 provided novel evidence for the theorized link between LC-NE activity and the exploration-exploitation tradeoff and revealed that increases in fluid intelligence were linearly related to increased pupil diameter during exploration—an exciting new finding that deserves further research. The proposed study will examine the relationship between the exploration-exploitation tradeoff and fluid intelligence in a controlled reinforcement learning environment. On day 1 participants will complete the Raven APM test to assess their fluid intelligence. On day 2 participants will perform a foraging task in which participants try to accumulate resources by exploring a foraging map. The foraging task will limit participants to two actions: exploring actions (moving to a new spot) and exploiting actions (collecting resources by staying on a located resource). In addition, the task will allow for active experimental manipulation of the strength of the exploration-exploitation tradeoff by manipulating the resource density of the map and starting position within the map. The map stimuli will be isoluminant and pupil foreshortening error will be corrected using the geometric model introduced in Chapter 5. The results of this study will serve as an important replication of Chapter 6 and will likely deepen our understanding of the interaction between fluid intelligence, the exploration-exploitation tradeoff, and the pupillary response and LC-NE system.
References


Cattell, R. B. (1971). *Abilities: Their structure, growth, and action*. Boston, MA:


Engle, R. W., & Kane, M. J. (2004). Executive attention, working memory capacity, and a two-factor theory of cognitive control. *The Psychology of Learning and
Motivation, 44, 145–199.


Van Der Meer, E., Beyer, R., Horn, J., Foth, M., Bornemann, B., Ries, J., . . .


A.1 SRSA technical details

As stated in Chapter 3, the successor representation was introduced to the reinforcement-learning literature by Dayan (1993) and was developed by White (1995). The SR is essentially identical to the fundamental matrix in the theory of Markov chains (Kemeny & Snell, 1976). More recently, Gershman, Moore, Todd, Norman, and Sederberg (2011) identified a formal connection between the SR and an influential model of episodic and semantic memory, the Temporal Context Model (e.g. Howard & Kahana, 2002; Sederberg, Howard, & Kahana, 2008).

We use a version of the successor representation that differs slightly from the standard definition (Dayan, 1993; White, 1995). The difference is that, when visiting a state $i$, our version does not include this same visit in the total (temporally discounted) number of visits to $i$. Assuming a first-order Markov chain with transition probability matrix $T$, our SR matrix $M$ is based on the power series:

$$M = T + \gamma T^2 + \gamma^2 T^3 + \ldots = T(I - \gamma T)^{-1}.$$  

The standard definition (Dayan, 1993; White, 1995) is based on the power series

$$I + \gamma T + \gamma^2 T^2 + \ldots = (I - \gamma T)^{-1}. $$

To revert to the standard formulation of the SR learning algorithm, the term $I_j$ in our Equation 2.2.1 must be replaced by $I_i$. In the special case when $\gamma = 0$, our algorithm tracks the transition matrix $T$ instead of the identity matrix $I$.

The proof that the temporal-difference learning algorithm in Equation 2.2.1 converges to the true successor representation $M$ (White, 1995) is a direct application of more general convergence proofs about TD($\lambda$) learning in the reinforcement-learning literature (Dayan, 1992; Jaakkola, Jordan, & Singh, 1994; Sutton, 1988). To ensure
convergence, it is necessary to decrease the learning rate $\alpha$ as the data accumulate.

The technical conditions include:

$$\sum_{n=0}^{\infty} \alpha_n = \infty \quad \text{and} \quad \sum_{n=0}^{\infty} \alpha_n^2 < \infty,$$

(A.1.2)

where $n$ is the number of observations (Dayan & Sejnowski, 1993, cited in White, 1995).

This indicates that the learning rate should be inversely related to the length of the data sequence. This in turn suggests a potential improvement of our eye-tracking analysis application. In the present article, we used a fixed $\alpha$ for all sequences regardless of length. It would be interesting to explore parameterizations that reduce the effective learning rate for longer sequences. The clipping of sequences longer than 100 fixations (described in the Method) is a crude way of regularizing the sequence length. Our present results indicate that, even with a fixed learning rate, the learning algorithm can accommodate substantial variability in length. As mentioned earlier, this is a major advantage over string-editing methods for comparing scanpaths. Varying the learning rate as a function of sequence length will provide additional robustness and reduce the variance of the estimates. This is a promising topic for future research.

Another promising possibility is to improve the feature-selection algorithm. Independent Component Analysis (ICA, Stone, 2004) may be better suited for eye-tracking applications than PCA because it relaxes the orthogonality constraint on the components. The SR matrices that correspond to psychologically relevant strategies are not necessarily orthogonal.

### B.1 Raven APM item relational scoring

Raven’s Advanced Progressive Matrices (APM) is traditionally scored as the total number of items correct. Preliminary SRSA analysis that used SR differences to
predict total number of APM items correct showed an overall trend for an increase in systematicity on session 2 (Hayes et al., 2011). To explore this finding in more detail, we needed to increase our overall power to resolve individual differences. The APM at its core tests the ability to extract relational information from complex, novel visual environments. Therefore, given that most APM items contain multiple distinct relations and an assortment of these relations are found within the 8 possible responses, we were able to increase our power to resolve individual differences by inferring the number of relations extracted for both correct and incorrect responses based on how many correct relations the chosen response contained.

When it was possible to tie the relation to a single feature, the feature to which the relational rule is applied is shown in parentheses (e.g., shape, shading, orientation, position, figure/ground, length). Finally, each of the eight possible responses were scored as either capturing (indicated by a 1) or failing to capture (indicated by a 0) each relation within an item. Four items (11, 14, 18, 27) had relations where partial credit was awarded for relational capture. For instance on item 11, response 1 was credited with .8 instead of 0 because it captured the addition relation but lacked a thin outside border around the figural item. For the other 24 remaining items, each response either clearly contained or lacked the relation(s). Table B.1 lists the relational score of each response for each item.
<table>
<thead>
<tr>
<th>Raven no.</th>
<th>Relations</th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>R4</th>
<th>R5</th>
<th>R6</th>
<th>R7</th>
<th>R8</th>
</tr>
</thead>
<tbody>
<tr>
<td>II-1</td>
<td>D3(shape)</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>D3(orient.)</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>CIR(lines)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>II-2</td>
<td>PP(position)</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>CIR(lines)</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>II-3</td>
<td>CIR(shape)</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>PP(position)</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>II-4</td>
<td>PP(shape1)</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>PP(shape2)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>II-5</td>
<td>PP(shade)</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>PP(shape)</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>II-6</td>
<td>PP(add)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>PP(subtract)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>II-7</td>
<td>ADD</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>II-9</td>
<td>ADD</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>II-10</td>
<td>PP(expand)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>PP(length)</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table B.1 *Raven Relational Scoring By Item.* Items are identified by their standard numbers in Raven’s Advanced Progressive Matrices. The Relations column lists which rules are present in the problem matrix and the feature to which that relational rule is applied in parentheses. The last eight columns represent the eight possible responses (moving left to right, top row is 1 2 3 4 and bottom row is 5 6 7 8) and whether or not they contain the corresponding relation.
Table B.1 continued

<table>
<thead>
<tr>
<th>Raven no.</th>
<th>Relations</th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>R4</th>
<th>R5</th>
<th>R6</th>
<th>R7</th>
<th>R8</th>
</tr>
</thead>
<tbody>
<tr>
<td>II-11</td>
<td>ADD</td>
<td>.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>II-12</td>
<td>SUBTRACT</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>II-13</td>
<td>D3(shape)</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>D3(orient.)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>CIR(lines)</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>II-14</td>
<td>PP(position)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.8</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>CIR(circle)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.8</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>II-15</td>
<td>ADD(figure)</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>ADD(ground)</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>II-16</td>
<td>SUBTRACT</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>II-17</td>
<td>D3(form)</td>
<td>0</td>
<td>0</td>
<td>.5</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>D3(shape)</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>II-18</td>
<td>D3(shape)</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>.2</td>
<td>1</td>
<td>.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>UNIQUE1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>UNIQUE2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>II-19</td>
<td>UNIQUE1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>UNIQUE2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>OPACITY1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>OPACITY2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>II-20</td>
<td>ADD</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>OPACITY</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>UNIQUE1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>UNIQUE2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Continued
<table>
<thead>
<tr>
<th>Raven no.</th>
<th>Relations</th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>R4</th>
<th>R5</th>
<th>R6</th>
<th>R7</th>
<th>R8</th>
</tr>
</thead>
<tbody>
<tr>
<td>II-21</td>
<td>D3(shade1)</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>D3(shape)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>PP(orient.)</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>PP(stretch)</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>D3(shade2)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>II-22</td>
<td>D2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>II-23</td>
<td>D2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>II-24</td>
<td>PP(wide lines)</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>PP(thin lines)</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>II-25</td>
<td>CIC(ground)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>CIR(figure)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>UNIQUE(shade)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>II-26</td>
<td>PP(orient.)</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>D3(shape)</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>II-27</td>
<td>D3(shape)</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>D3(form)</td>
<td>0</td>
<td>0</td>
<td>.3</td>
<td>0</td>
<td>.3</td>
<td>0</td>
<td>1</td>
<td>.3</td>
</tr>
<tr>
<td>II-28</td>
<td>D3(shape1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>D3(shape2)</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>D3(number1)</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>D3(number2)</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>II-29</td>
<td>D3(shape)</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>PP(orient.)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>D3(length)</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>