ABSTRACT

STOKES, THOMAS ALLEN. Human-Graph Interaction: User, Task, and Display Related Factors' Effect on Performance and Search. (Under the direction of Douglas J. Gillan).

Information visualization techniques have become a popular way to explore large scientific datasets. However the rate of research on graph reading has not kept up with the rate of development of new visualization techniques. This has left us with holes in our ability to assess the usefulness and usability of graphs and information visualizations. The experiment presented in this paper attempts to move towards a more complete model of graph reading, which may later be used to develop usability assessment tools and design guidelines for all forms of information visualization.

© Copyright 2017 by Thomas Allen Stokes

All Rights Reserved

Human-Graph Interaction: User, Task, and Display Related Factors' Effect on Performance and Search

by
Thomas Allen Stokes

A thesis submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Master of Science

Psychology

Raleigh, North Carolina

2017

APPROVED BY:

_____	_____
Jing Feng	Christopher B. Mayhorn

_____
Douglas Gillan
Committee Chair

BIOGRAPHY

Thomas Allen Stokes was born in Raleigh, North Carolina where he was raised by his parents, Michael T. Stokes and Sabrina M. Stokes. Thomas received a B.A. in Psychology in 2013, and enrolled in the Human Factors and Applied Cognition program the following fall, working under the guidance of Dr. Doug Gillan.

ACKNOWLEDGMENTS

I'd like to start by thanking my parents, Michael and Sabrina Stokes. Without their support through 26 years I would not be who I am today.

To my advisor and committee chair, Dr. Gillan, you welcomed me into your lab and gave me the opportunity to pursue graduate studies—you've been a wonderful mentor and I owe you many thanks. You instilled in me the confidence to pursue and complete this experiment.

Additionally I'd like to thank my committee members, Dr. Feng and Dr. Mayhorn, who provided guidance and valuable insights throughout this experience.

I would be remiss if I did not acknowledge the members of my cohort, who have become some of my closest friends (Caleb Furlough, Nick Mudrick, Lawton Pybus, Michelle Taub, Allaire Welk, and Olga Zielinska)—I couldn't imagine experiencing graduate school with a better, brighter group than you all. Friendship is a valuable thing, and I lucked into a group of 7 incredible individuals.

Finally, Paulina (and Denver) thanks for always being there for me through difficult times, and celebrating the good ones. Your patience, and dedication to me no-doubt helped make this possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Advances in technology have brought about increasingly powerful computers with better visual displays and a rise in the amount of data that can be processed and stored (Borwein, & Bailey, 2015; Dipert, 2010). This coupling of greater amounts of data and more powerful computers with detailed displays has facilitated rapid development in the field of information visualization. "Visual representations and interaction techniques take advantage of the human eye's broad bandwidth pathway into the mind to allow user to see, explore, and understand large amounts of information at once" (Thomas & Cook, 2005, p. 30). In this way, we understand the "why" behind information visualization; it's a simple yet elegant solution to the problem of having to acquire insight from mountains of data. However, the processes that underlie the way people gain information from these visual representations have not been studied at the same rate as the advancement of our available visualization techniques; so, researchers are catching up on the "how" of user's interactions with information visualization. Unfortunately, this lack of knowledge about human-graph interactions has left us with gaps in our ability to adequately assess the usability and effectiveness of visualizations (Thomas & Cook, 2006).

The vast majority of research so far in information visualization has fallen into four broad categories (Liu, Cui, Wu, & Liu, 2014), one of which is empirical research into the theoretical foundations of graphs through experiments and usability studies. Overall, the literature shows that three main components play a role in human-graph interactions; user, task, and display related factors (e.g., Gillan, 2009).

One example of the empirical approach is from Lohse who proposed a cognitive model derived from recording eye tracking when people read simple graphs (1991), then performed experiments to validate his theory (1993). In these experiments, task demands and display types varied, and were identified as being important factors in predicting performance. Lohse (1991) identified search patterns as a meaningful measure for insight into the subtasks that a graph reader performs while interacting with a graph, but did not analyze differences in search patterns that may be associated with user, task, or display differences. Rather, Loshe's model only considered one possible search pattern, termed the semantic trace, which assumes a predictable sequence of fixations based on the graph reader's task.

The idea of one search pattern would appear to conflict with research on individual differences. For example, research on expertise has shown that domain experts spend their time more efficiently, in terms of total time and time on target information, when engaged in search than novices (Hirsch, 1997; Kuhlthau, 1999; Marchionini, 1995; McDonald & Stevenson, 1998; Patel, Drury, & Shalin, 1998). Additionally, experts make faster relevancy judgments about the information they encounter (Marichini et al, 1993), and organize their information into meaningful patterns (Chase & Simon, 1973). Specifically, in situations of graph reading, domain experts have superior performance compared to novices when looking at graphs with information pertaining to their field. Novices tend to rely on low-level processing and do not differentiate meaningful from nonmeaningful details, whereas, experts see just as many non-meaningful relationships but, due to their expertise, are better able to

identify them and instead direct their focus primarily to meaningful trends (Freedman & Shah, 2001; Gattis & Holyoak, 1995). Previous research shows that the outcomes of expert vs. novice search has measurably different outcomes, so it stands to reason that the process of search patterns may also show differences. Considering this, models of graph reading should include search pattern as a variable that affects performance, but also consider user-related characteristics that feed into search.

Gillan and many colleagues published a series of papers on componential models of human-graph interaction (Gillan & Neary, 1992; Gillan & LaSalle, 1994; Gillan & Lewis, 1994; Gillan & Harrison, 1998; Gillan, 2000; Gillan & Callahan, 2000; Gillan, 2009). Based on a preliminary task analysis, Gillan & Lewis (1994) developed the Mixed Arithmetic Perceptual (MA-P) model, which proposed five main components of human graph interactions -- searching for each indicator named in a question, encoding components for each indicator, arithmetic operations, spatial comparisons, and response. It is also important to note that the model does not indicate a single order in which the operations are performed, rather order is a function of task, graph and user factors. This MA-P model evolved over the series of experiments and in summary shows a model of human-graph interactions where task, user, and display-related factors all influence performance on analytical activities. This series of experiments also show interactions between user and task factors, and task and display factors.

In addition to culminating in a fairly comprehensive model, the papers by Gillan and colleagues introduced a potentially pivotal idea to the human-graph interaction literature:

visualization ability, which is strongly related to mental imagery (for further reading see: Shepard & Metzler, 1971; Kosslyn, 1975; Kosslyn, Ball, and Reiser, 1978), plays an important role in graph reading. A primary point in the discussions of the publications in this series is that graphs, as spatial representations of data, require a user to manipulate stimuli mentally in order to perform analysis and comparisons. Gillan (2009) also proposed that, although some comparisons can be between presented stimuli, some others may require the graph reader to use mental anchors or prototypical stimuli they mentally envision. Despite Gillan's focus on explaining performance through visualization ability, no specific follow-on research using visualization ability as a user-related predictor for performance has yet been done. There are few studies like the one by Gillan and Callahan on pie chart segments (2000) whose results focus on perceptual and cognitive imagery explanations. However, evidence suggests that low-level analytical activities involving graph reading likely involve translating, moving, rotating, or comparing of mental images. If differences in visualization ability affect graph reading performance, it would be a valuable factor to include in a more complete model of human-graph interaction.

Despite the potential applications and ability to explain performance differences, research into human-graph interactions rarely includes process measures or eye tracking (although see Goldberg & Helfman, 2011; Huang, 2007; Peebles & Cheng, 2003), and has never to this point looked into user, task and display factors as predictors for search pattern differences. Additionally limited effort has been made to synthesize the human-graph interaction literature in a way that combines user-related (specifically regarding expertise and

visualization ability), task-related, display-related factors, and their associated interactions, as well as how they impact both search pattern and task performance in terms of response time and accuracy. In addition to increasing the ability to explain performance outcomes, a comprehensive model of human-graph interaction can aid in the design of future data visualizations and graphs. If designers and information-graphics makers can somehow understand not only the quantitative outcomes but also the qualitative process differences, including search patterns and strategies, then techniques can be designed around a more complete model of human-graph interaction. As a consequence of the application of the more complete model, both the utility and usability of graphics can be improved.

The current research set out to study how user, task, and display related factors affect search pattern and overall performance on basic analytic tasks with graphs. Search process (eye tracking) data was used to find potentially different strategies associated with the interaction of these factors, and aid our understanding of exactly what is done when performing basic analytic activity, providing a base for future development of research into the same paradigm.

The hypotheses included:

I. Search patterns will differ based on user, task, and display related factors.

II. Visualization ability (a user-related variable) will moderate the relation between task type and performance (response time and accuracy), leading to a significant interaction.

III. Expertise (a user-related factor) will affect performance (response time and accuracy) and search pattern type in domain-specific tasks.

IV. Performance (reaction time and accuracy) will be improved by use of data point labels (i.e., performance with "grables" (Hink, Eustace, Wogalter, 1998) will be better than performance with regular graphs).

Figure 1 shows the combination of these hypotheses. To assess the model, the researchers performed an experiment where display and task factors were manipulated, and user factors were measured by instruments.

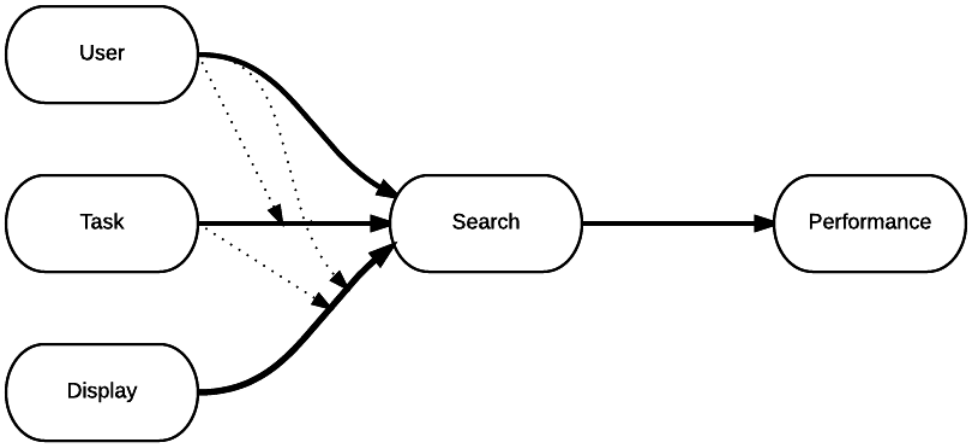


Figure 1. Graphical representation of proposed model of human-graph interaction.

As indicated previously, the gaps in the literature point towards visualization ability and domain expertise as important user-related factors of interest. Both of these were measured by questionnaires. Domain expertise was assessed by an instrument developed by the researchers. Visualization ability was measured by the Object-Spatial Imagery Questionnaire (OSIQ) (Blajenkova et al., 2006), a validated two-factor measure of

visualization ability. For further information about the scale's development, reliability, and validity see Appendix D.

Task type was manipulated so that there were analytical, holistic and domain-specific tasks. Previous research (see Amar, Eagan, & Stasko, 2005) has shown that there are many ways of categorizing low-level analytic activity, but the way that made the most sense for testing the proposed hypotheses was to include thee general categories – (1) computational trials in which participant have to use graphs as the data-source for computations, (2) holistic tasks that afford the participant the ability to perceive information either serially or holistically, and (3) domain-specific tasks in which a domain expert would be able to interpret information from the graph and project their understanding of the domain into their responses.

The present research used spider graphs -- a common type of object graph that shows both symmetry of values and their magnitudes (Goldberg & Helfman, 2010) -- that are similar to star graphs. Figure 2 provides an example of a typical spider graph. Both spider and star graphs consist of lines radiating from a central point, with the amounts of different variables represented by the lengths from the central point to marks on the different lines (spider graphs) or the lengths of the different lines (star graphs). These types of graphs have been used in past experiments where results suggested evidence of holistic vs. analytical processing (Gillan & Harrison, 1999). The use of object family graphs for this study is important in that they may permit the observation of differences in analytic vs. holistic processing as a function of visualization ability and expertise.

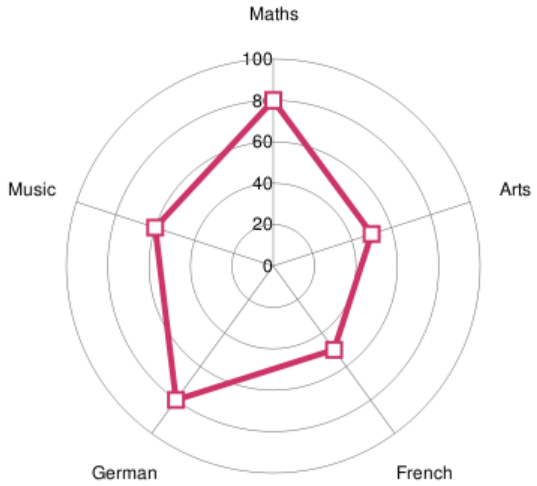


Figure 2a. A typical representation of a spider graph.

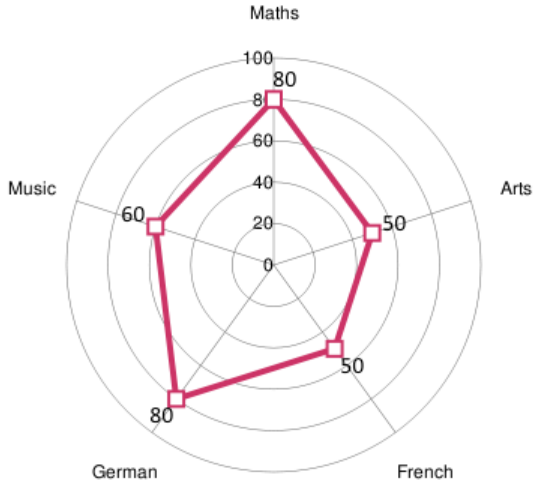


Figure 2b. A typical representation of a spider grable.

In addition, researchers examined the effect of data point labels, which provide the numerical values of each indicator on the graph. Previous research has shown the performance differences between tables, graphs, and "grables" (for example: Hink, Wogalter, Eustace, 1996; Hink, Eustace, & Wogalter, 1998; Meyer, Shinar, & Leiser, 1997), but no research has

been done while also observing search pattern and including the user-related factors accounted for here.

METHOD

Research Design

The research presented in this paper was conducted in a single experiment with display and task variables manipulated as within-subject variables. Task type and display type were manipulated as follows:

- Task type (3 levels) – within subject

    1. Computational
    2. Holistic
    3. Domain Specific

- Display type (2) - within subject

    1. Regular Graphs (no data point labels)
    2. Grables (graphs with data point labels)

Additionally, the following user characteristics were measured (but not manipulated) through questionnaires:

- Domain knowledge - Football interest and expertise questionnaire
- Visualization ability - OSIQ

For the experimental task, participants performed low-level data analysis tasks using sets of four spider graphs. There were three basic analysis/task types: a computational task (computing an average), a holistic task (identifying which two spider graphs were most similar to one another), and a domain specific task (choosing an optimal solution or strategy for a situation). The data presented to participants was a set of static, small multiple displays

of spider graphs showing statistics for four NFL quarterbacks. The quarterbacking statistics presented in the spider graphs included completion percentage, yards, touchdowns, interceptions, and total quarterback rating all on individual spokes at five different angles radiating from a fixed center point; for an example of a typical trial see Appendix A. There were two different display types: (1) Grables contained data point labels, which provide numerical values for each indicator and (2) a second type without any numerical labels for the indicators. The reason for using these data point labels was to examine user reliance on exact values, when directly presented, and whether or not this affects subsequent search and performance. Search patterns were recorded for these tasks by an eye tracker, with gaze/search pattern and fixation length information capabilities. Performance was also measured in terms of accuracy and response time.

In addition to the experimental tasks this experiment assessed the results from three questionnaires. These included one questionnaire for basic demographic data, another for assessing overall football experience and knowledge level (so that subjects could be categorized generally as novice or experienced in regards to the data presented to them), and the OSIQ for assessing visualization ability using an object and a spatial factor (Blajenkova et al., 2006)

Participants

Forty-eight participants were recruited from introduction to psychology courses at North Carolina State University as a part of course research credit using the online recruitment and scheduling system, Experimetrix. A sample size of 46 was determined to

have sufficient power, from a calculation to find a medium size effect (Cohen's $F=.25$) with Power=.95, at the $\alpha=.05$ significance level—the obtained sample exceeds the target sample size by 2 participants. Participants were required to be at least 18 years old and have normal or corrected-to-normal vision with either naked eyes or contact lenses. Participants needing eyeglasses were not included in the study because the glare on the lenses of the glasses is a known cause of errors in recording eye tracking due to glare on the glass lenses.

Materials

Lab Setting

The experiment took place in a well-lit research lab, with closed doors and as little noise exposure as possible. Participants were run one at a time, and the only two people in the room at any time were the researcher and the participant.

Computer and Eye Tracking Station

The experimental set-up included a Dell dimension series 4500, with Windows XP, using an ASUS 17" monitor with 1024 x 768 resolution, a standard QWERTY keyboard, and trackball mouse. Connected to this computer was an EyeTrac 6 system from Applied Science Laboratories. The eye tracker was calibrated, for each participant, using standard 9-dot-coordinate calibration with automatic panning. Participants sat in a four-legged, non-swivel chair, approximately 22 inches away from the monitor.

Stimuli

Experimental stimuli consisted of small multiples of spider graphs made in Microsoft Excel's built-in chart making tools. For an example of what a typical stimulus set looked

like, see Figure 3. In total, there were 18 experimental trials -- 6 different sets of graphs (3 including data point labels and 3 without), with 3 different questions for each graph (one computational, one holistic, and one domain-specific).

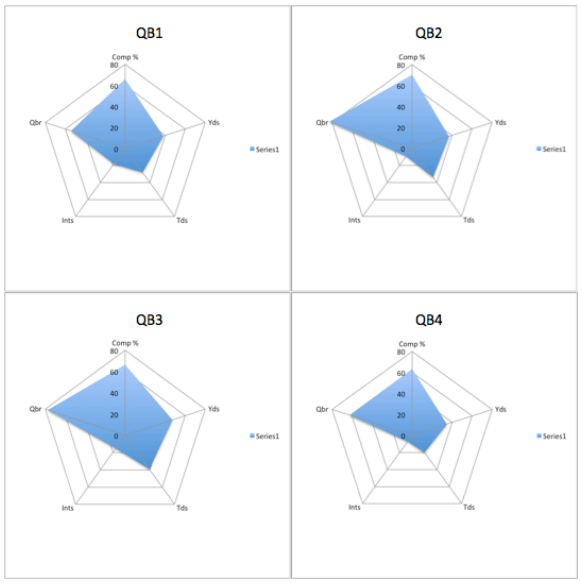


Figure 3. Example trial stimuli.

The stimuli and associated low-level analysis questions were presented using a program made specifically for this study and delivered though Qualtrics' online survey-making tool (Qualtrics, 2017). The electronic survey tool had the capability to randomly present elements, counterbalance data set order, and record participant responses as well as response times and accuracy.

Demographics Questionnaire

This questionnaire recorded basic participant demographics of age, gender, and college major.

Arithmetic Ability Quiz

Participants completed a set of 10 basic arithmetic questions—these items can be found in Appendix E. This instrument was included to get a baseline of the participants' general mathematical abilities. Participants were given 5 minutes to complete the set of questions, and were not allowed any calculating devices. Score on the quiz was equal to the number of items answered correctly.

Football Interest and Expertise Questionnaire

This questionnaire assessed participants' experience, knowledge base, and level of interest in football. The experimenters developed this scale for use in this study and consist of questions thought to be indicators of football knowledge (see Appendix B). In addition to the Likert items participants were asked to estimate the number of hours per week they spend in football related activities (e.g. playing fantasy, watching, reading articles). This resulted in two data points representing expertise for each participant:

1. raw number of hours per week spent in domain-related activities
2. average Likert item response

OSIQ

The OSIQ (Object-Spatial Imagery Questionnaire) was administered to assess participants' visualization ability on two factors. The two factors of the scale measured two different aspects of mental imagery; one for the schematic spatial operations (i.e. scanning, rotation, scaling) named the "spatial" scale, and one for the vividness of imagery named the "object" scale (Blajenkova et al., 2006). Scores for each subscale are calculated by

averaging the response for each Likert item associated with that subscale (with negatively worded item responses reverse coded). This calculation results in a minimum score of 1 and maximum of 5 for each subscale.

Procedure

When the participant arrived at the laboratory, the experimenter confirmed with them that they had normal or corrected-to-normal vision and were not wearing eyeglasses (for previously mentioned eye-tracking reasons). Next, participants were instructed to read, date, and sign paper copies of the consent form. After completing the consent forms participants filled out the basic demographic survey followed by the arithmetic assessment.

Once the demographic survey and arithmetic assessment were completed, the experimenter calibrated the eye tracking system using a standard 9-dot calibration technique. The distance between the participant and the display was approximately 22 inches.

After calibration, the experiment began. First, there were three practice trials to acquaint the participants with the tasks they would be asked to perform in the experiment. The subject matter for the practice trails (the number and type of animals on different farms) was different from that in the actual trials (NFL passing statistics) so that participants would not be exposed to the experimental stimuli. Following the practice trials, the participants received the 18 experimental trials in random order.

After completion of the experiment, the experimenter administered the OSIQ, and football interest and expertise questionnaire in counterbalanced order. Once data collection

was complete, the participant was debriefed, given an opportunity to ask the researcher questions regarding the nature of the experiment, and given credit for participating.

RESULTS

Data Preparation

Prior to hypothesis testing questionnaire and survey responses were scored, then descriptive statistics and distributions were obtained for each. A summary of these variables grouped by characteristic measured and the correlations between the measures can be found in Table 1 and Table 2 below. Measures were calculated as described in the method. As one can see in Table 2 there is a significant relationship between arithmetic ability and spatial visualization ability; their relationship and its impact on subsequent analyses is assessed in Appendix F.

Table 1. Descriptive Statistics for User Variables

Characteristic	Mean	St.Dev	Median	Skew	Kurtosis	Min	Max
Visualization Ability							
Spatial score	3.28	.58	3.33	-.63	.63	1.67	4.40
Object score	3.23	.51	3.27	-1.10	1.43	1.60	4.00
Combined Score	3.26	.34	3.28	-.28	.23	2.40	3.93
Expertise							
Survey Reponses	3.07	1.13	3.15	-.25	-.83	1.00	5.00
Hours per week	4.41	4.67	3.25	2.02	5.83	0	24.00
Log-hours per week	.49	.40	.06	.14	-1.07	0	1.38
Mathematical ability							
Number questions answered correctly	8.42	1.54	9.00	-.67	-.76	5.00	10.00

Table 2. Correlations Between User Variables

Measure	1	2	3	4	5	6
1. Expertise-Survey	--					
2. Expertise-Hours	.766**	--				
3. Math # Correct	-.050	-.095	--			
4. Object Score	-.001	.024	-.069	--		
5. Spatial Score	-.146	-.182	.459**	-.245	--	
6. Combined Score	-.126	-.139	.343*	.551**	.674**	--

Note: ** indicates p< .01, * indicates p<.05

For subsequent F-family tests involving nominally grouped task and display factors, the user characteristics must also be split into levels. Since a primary goal of the research was to observe individual differences in high verses low levels of expertise and high verses low visualization ability the variables were split into three equal groups with cutoff points at the 33rd and 66th percentile. This resulted in three equal groups for each variable; low, medium and high for both visualization ability and expertise scores. This method was chosen as superior to a median split as it separates out high and low scoring individuals from the individuals who score in a "typical" range—which more-so fits the goals of the research.

Hypothesis I: Search Patterns Differ by User, Task & Display

Testing the first hypothesis of this study (search patterns will differ based on user, task, and display related factors) required analysis of the process data from eye tracking. Gillan and Cooke (2001) have proposed PRONET as a method for analyzing sequences of process data by using Pathfinder networks (see Schvaneveldt, 1990). To apply this method to the current study, the researchers identified each meaningful area of interest (AOIs), and

labeled them; the figure below highlights the areas of interest over a typical trial. These areas were the same across all trails and included all four graphs, the question text, the response area, and the button to proceed to the next trial.

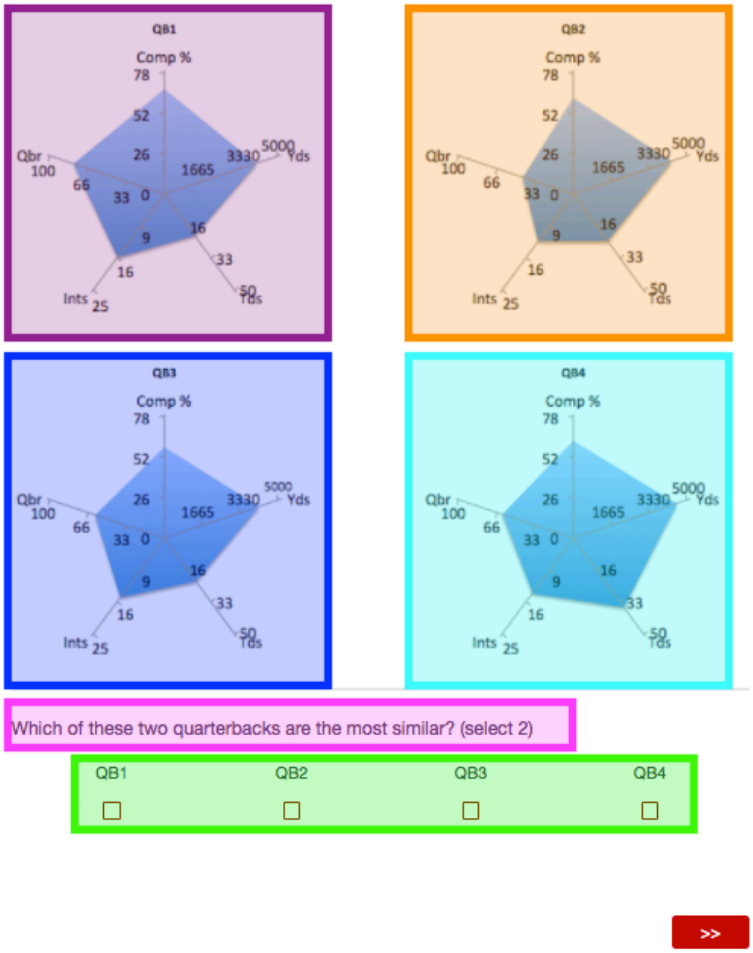


Figure 4. AOIs for Trial Stimuli.

Next, using R the following steps were followed:

1. Using the known area ranges (for an AOI) each fixation from eye tracking data was coded into its corresponding area of interest

2. Transitions from a fixation at one AOI to another were recorded.

3. These transitions were then organized into a matrix, where the number in one cell represents the number of transitions from the AOI in the column to the AOI in the row.

4. The transition frequencies were converted into transition probabilities by dividing the tally in each cell by the total for the column

It is important to note that the conditional transition probability from 'AOI alpha' to 'AOI beta' would be different than that of 'AOI beta' to 'AOI alpha'. This means that the transition probabilities are unbalanced data and require a full matrix, rather than a half matrix.

Using this matrix of transitional probabilities, Pathfinder networks were obtained, with the assistance of j-Path, a program that provides a graphical user interface to obtain and compare pathfinder networks (Interlink Inc, 2017). Networks were obtained for a group on a given task (e.g. low spatial visualizers on domain specific tasks)—in other words the obtained networks represent the mean latent scan pattern for the whole group for a given task type.

Once Pathfinder networks were obtained, they were compared using the coherence (C) and similarity (S) metrics. C quantifies the consistency or reliability of the data in a pathfinder network (*cf* Cronbach's alpha). C has a maximum value of 1, and scores below .15 are generally regarded as being extremely low (Interlink Inc, 2017). S is a metric that quantifies the similarity between two pathfinder networks, and be seen as very similar to a

correlation, however instead of ranging from -1 to 1 its values can range from 0 to 1, with higher values indicating greater similarity. Additionally, S has a significance value associated with it that is calculated from the hypergeometric distribution using the number of links in each network being compared and the number of shared links between them. The significance value reported is the probability of having a greater number of shared links than the observed number. Values under .05 indicate significant similarity, and numbers over .95 indicate significant dissimilarity. Tables 3a, 3b, 4a, 4b, 5a, 5b, 6a, 6b, 7a, 7b, 8a, 8b, 9a, 9b, 10a, and 10b show the results of PRONET analyses for expertise (as measured by hours on task, and survey responses), spatial visualization ability, and object visualization ability.

PRONET: Expertise and Search

In terms of expertise, regardless of whether it was measured by a questionnaire or number of hours of weekly experience, there was no significant difference in the networks on holistic or computational type trials (see Tables 3a and 5a). However, different levels of expertise had statistically similar networks for domain-specific tasks; a finding that directly conflicted with the hypotheses of this experiment as it was expected that domain expertise would guide a person to a different way of looking through the graphs. PRONET networks and heatmaps for experts and non-experts on domain specific trials are shown in figures 5a, 5b, 5c, and 5d. The heatmaps overlay red coloring over the area of the stimuli that was viewed most (with dark red indicating greater viewing time). When looking at these pairs of heatmaps one cannot discern a meaningful pattern from them. The PRONET networks show the latent scanning pattern for the given group, with links between commonly associated

AOIs indicated by arrows; again one cannot discern any meaningful differences between the pairs of PRONET networks.

The obtained networks for experts had higher coherence (see Tables 4a and 6a) than those of non-experts (i.e. expert scan patterns were more consistent or more predictable than non-experts), especially for domain-specific tasks. Taken together these findings show that while experts and non-experts scanned-through the graphs very similarly, the experts were far more consistent in their overall pattern and one has a greater ability to predict the next fixation for experts than novices, especially when the task at hand relies on the expert's domain knowledge.

Table 3A. Similarity: Expertise-Hours, Low vs High- Graphs

	Task Type		
Expertise-Hours	Computational	Holistic	Domain-Specific
S (sig)	.27 (.574)	.33 (.105)	.46 **(.017)**[*]

Note. [*] Indicates significant similarity. [†] Indicates significant difference.

Table 3B. Similarity: Expertise-Hours, Low vs High- Grables

	Task Type		
Expertise-Hours	Computational	Holistic	Domain-Specific
S (sig)	.59 **(<.001)**[*]	.49 (.164)	.70 **(<.001)**[*]

Note. [*] Indicates significant similarity. [†] Indicates significant difference

Table 4A. ΔCoherence Low-to-High: Expertise-Hours- Graphs

	Task Type		
Expertise-Hours	Computational	Holistic	Domain-Specific
ΔC Low to High	.77	-.04	1.24

Note. Coherence does not have a significance value.

Table 4B. ΔCoherence Low-to-High: Expertise-Hours- Grables

	Task Type		
Expertise-Hours	Computational	Holistic	Domain-Specific
ΔC Low to High	-.11	.40	1.52

Note. Coherence does not have a significance value.

Table 5A. Similarity: Expertise-Survey, Low vs High- Graphs

	Task Type		
Expertise-Survey	Computational	Holistic	Domain-Specific
S (sig)	.20 (.748)	.45 (.173)	.47 **(.010)**[*]

Note. [*] Indicates significant similarity. [†] Indicates significant difference.

Table 5B. Similarity: Expertise-Survey, Low vs High- Grables

	Task Type		
Expertise-Survey	Computational	Holistic	Domain-Specific
S (sig)	.56 **(<.001)**[*]	.50 **(.015)**[*]	.40 (.055)

Note. [*] Indicates significant similarity. [†] Indicates significant difference.

Table 6A. ΔCoherence Low-to-High: Expertise-Survey- Graphs

	Task Type		
Expertise-Survey	Computational	Holistic	Domain-Specific
ΔC Low to High	.38	.25	.84

Note. Coherence does not have a significance value.

Table 6B. ΔCoherence Low-to-High: Expertise-Survey- Grables

	Task Type		
Expertise-Survey	Computational	Holistic	Domain-Specific
ΔC Low to High	-.12	.33	1.16

Note. Coherence does not have a significance value.

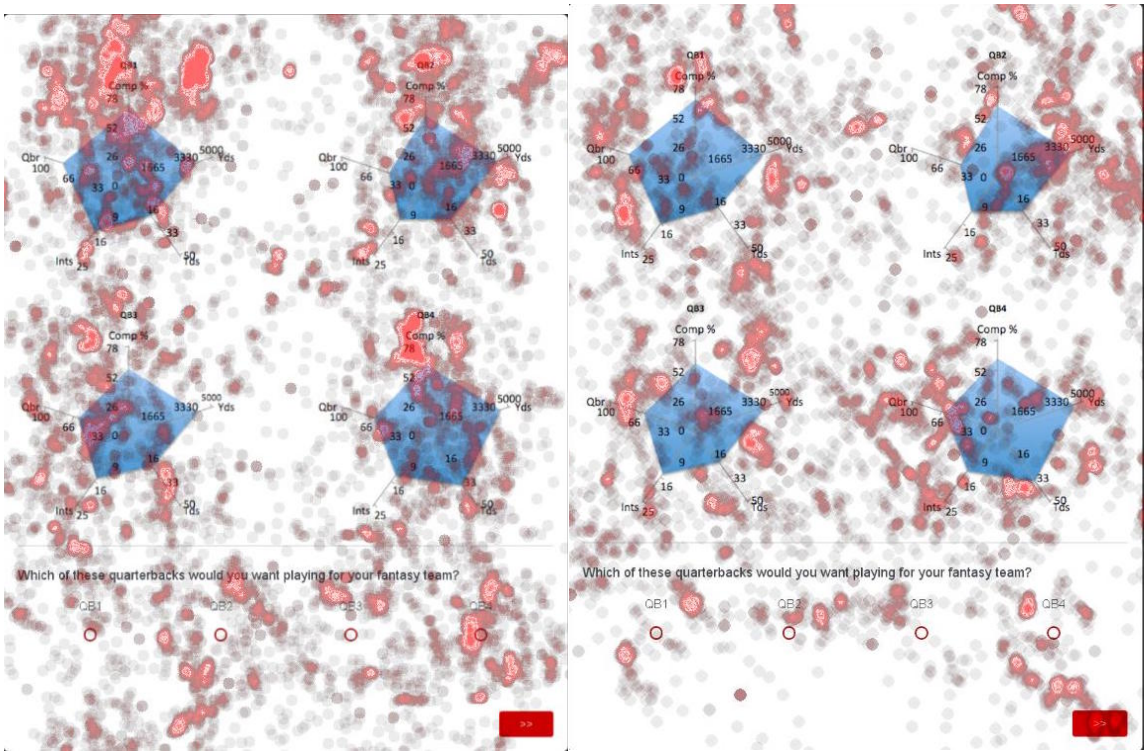


Figure 5a. Comparison of High (left) and Low (right) Expertise (measured by hours) in Domain Specific task-Heatmap

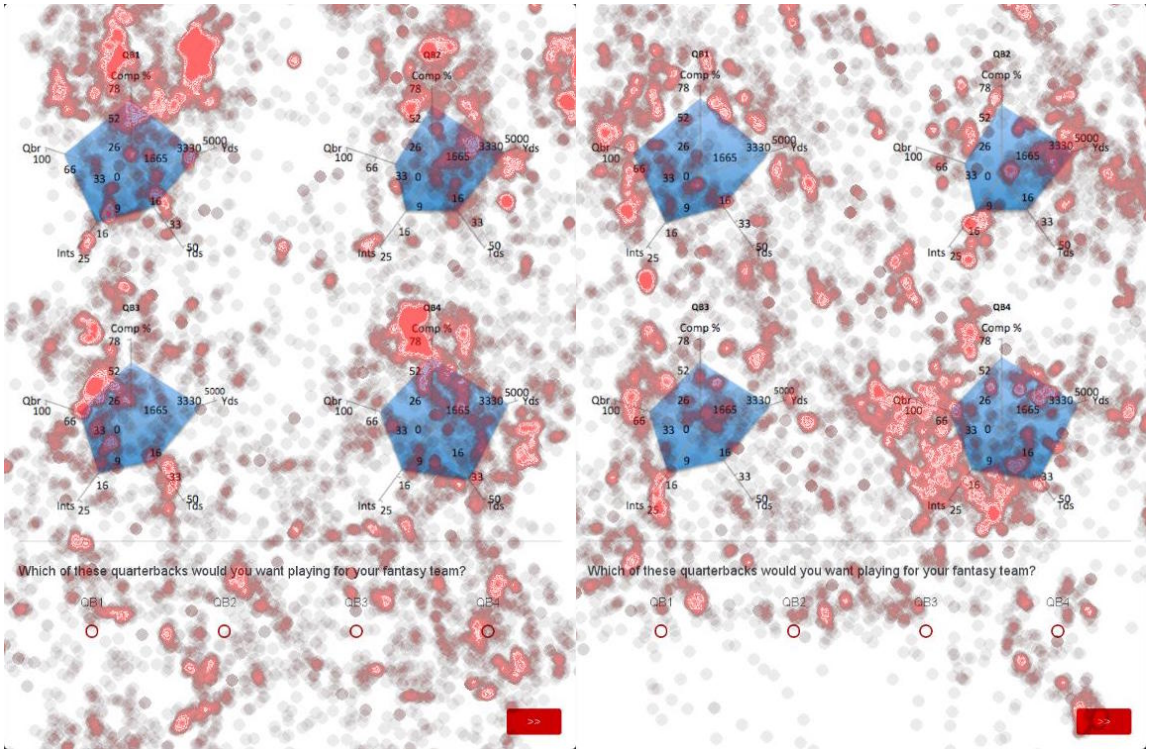


Figure 5b. Comparison of High (left) and Low (right) Expertise (measured by survey items) in Domain Specific task- Heatmap

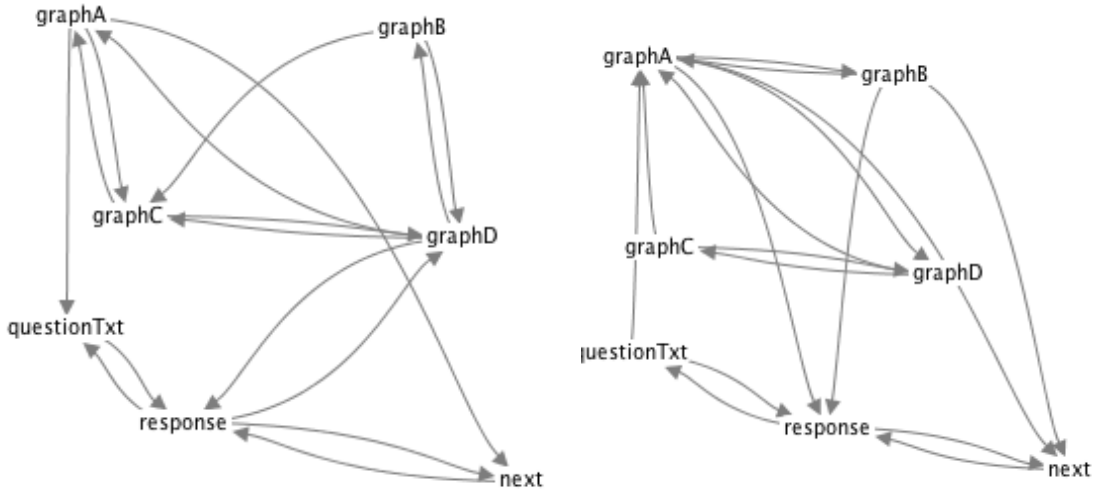


Figure 5c. Comparison of High (left) and Low (right) Expertise (measured by reported number of hours) in Domain Specific task- Pathfinder Networks

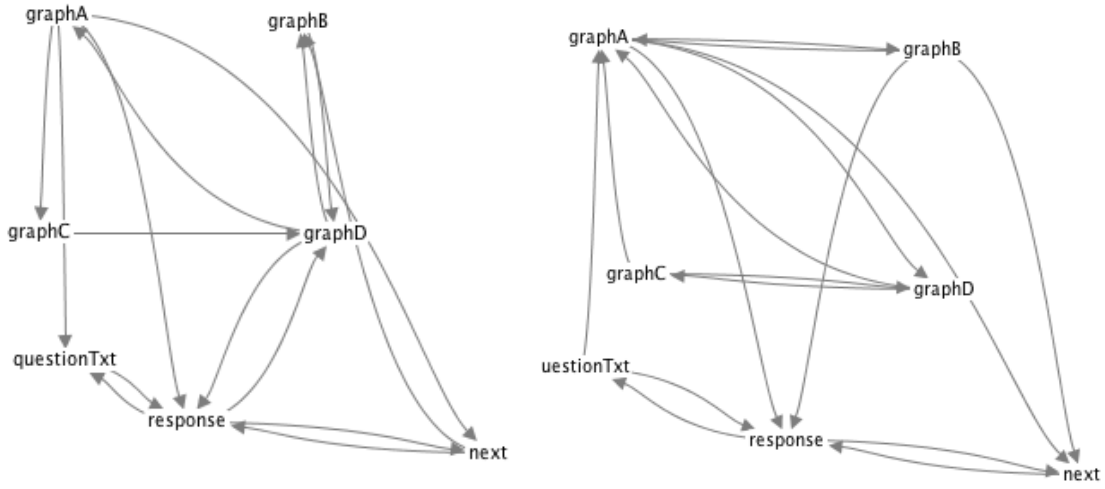


Figure 5d. Comparison of High (left) and Low (right) Expertise (measured by survey items) in Domain Specific task- Pathfinder Networks

PRONET: Visualization Ability and Search

For holistic and domain-specific tasks spatial visualization ability yielded no significant differences or similarities in pathfinder networks, however there was a significant difference for computational trials (see Table 7a). Additionally, for computational tasks, those who score low on spatial visualization ability have higher coherence than those who score highly (see Table 8a). This finding suggests that the higher one scores on spatial visualization ability the less likely they are to follow a particular scan pattern and will be less consistent in their scan pattern for tasks that involve computations. Heatmaps of fixations can be seen in Figure 6a, where one can see that high spatial visualizers may have spent more time responding to questions by the increased focus on the answer box. The corresponding pathfinder networks are shown in Figure 6b.

Adding labels to graphs significantly affects these trends. First of all, grables increase high spatial visualizers' coherence for computational tasks, but decreases for holistic and

domain specific tasks (see Table 8b as compared to 8a). Additionally, grables make it so that both low and high spatial visualizers have significantly similar scan patterns across all observed task types.

Table 7A. Similarity: Spatial Visualization Ability, Low vs High- Graphs

	Task Type		
Spatial Vis Ability	Computational	Holistic	Domain-Specific
S (sig)	.09 **(.968)** †	.41 (.105)	.28 (.783)

Note. * Indicates significant similarity. † Indicates significant difference.

Table 7B. Similarity: Spatial Visualization Ability, Low vs High- Grables

	Task Type		
Spatial Vis Ability	Computational	Holistic	Domain-Specific
S (sig)	.43 **(.021)** *	.45 **(.015)** *	.52 **(<.001)** *

Note. * Indicates significant similarity. † Indicates significant difference.

Table 8A. ΔCoherence Low-to-High: Spatial Visualization Ability- Graphs

	Task Type		
Spatial Vis Ability	Computational	Holistic	Domain-Specific
ΔC Low to High	-.84	.61	.62

Note. Coherence does not have a significance value.

Table 8B. ΔCoherence Low-to-High: Spatial Visualization Ability- Grables

	Task Type		
Spatial Vis Ability	Computational	Holistic	Domain-Specific
ΔC Low to High	-.45	-.39	.09

Note. Coherence does not have a significance value.

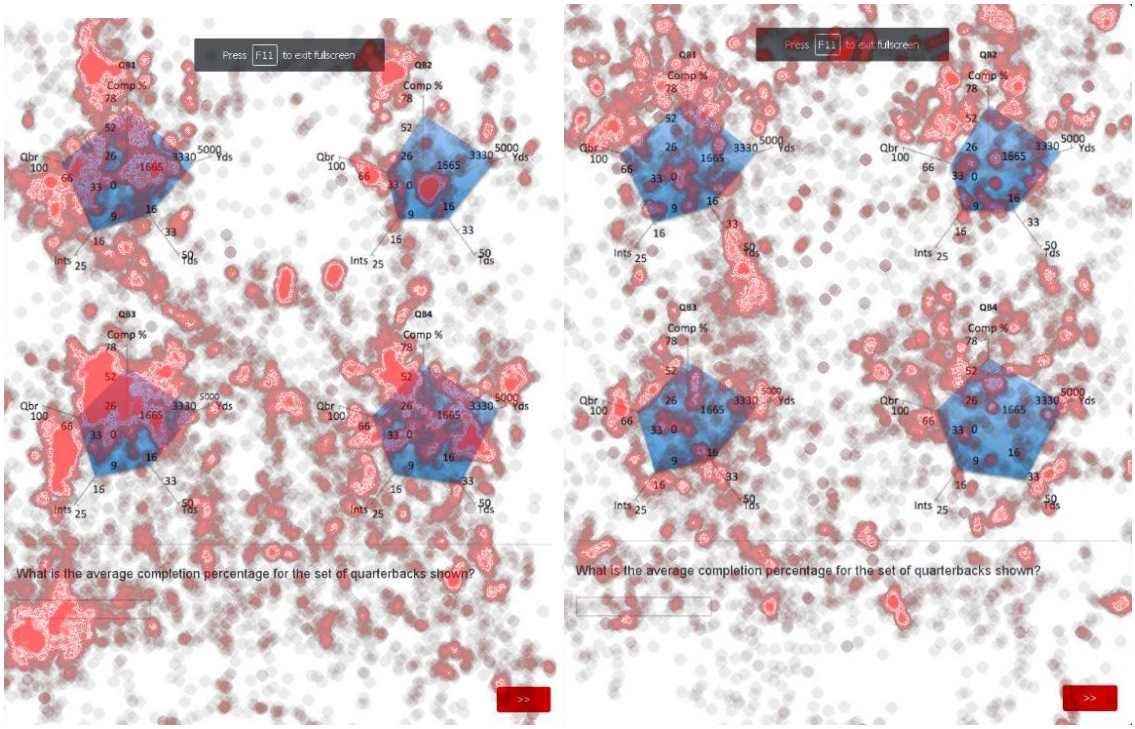


Figure 6a. Comparison of High (left) Spatial and Low (right) Spatial Visualization Ability on Computational Task- Heatmaps

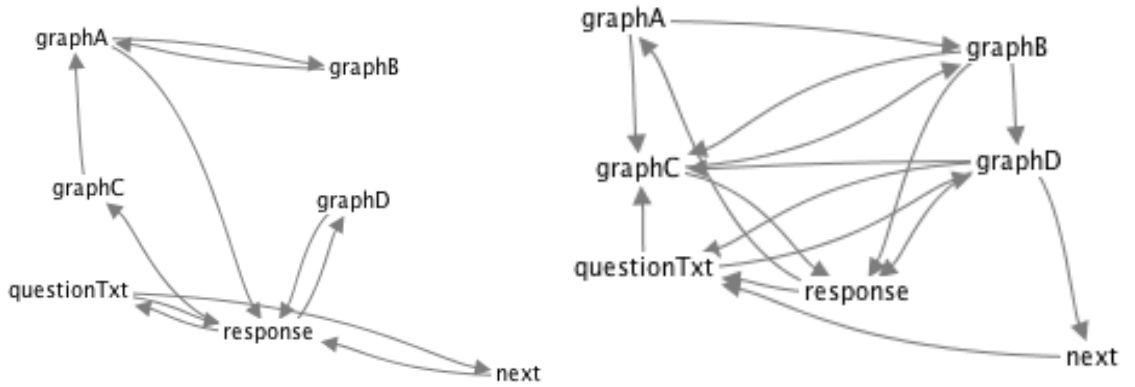


Figure 6b. Comparison of High (left) Spatial and Low (right) Spatial Visualization Ability on Computational Task- Pathfinder Networks

Object visualization ability does not seem to account for any scan pattern differences in either computational or domain specific tasks, however significant similarity between the networks for low and high scores was found for holistic tasks (see Table 9a). Heatmaps for

these trails are shown in Figure 7a, and the corresponding pathfinder networks are shown in Figure 7b; readers should note no major, noticeable differences in either of these visuals.

While low and high object visualizers scanned similarly, it appears that high scores on object visualization ability lead to higher coherence scores for holistic trials, and low scores lead to higher coherence on computational trials, which matched the pattern seen in spatial visualization ability (see Table 10a).

Again, grables changed the patterns of results for object visualization ability (see Table 10b as compared to 10a). First off, the graphs with labels yielded higher coherence for the high object visualization group when doing computational tasks, but lowered their coherence for the holistic and domain specific tasks. Additionally, high and low object visualizers had statistically similar PRONET networks on computational tasks and domain specific tasks, where without grables they were only statistically similar for holistic tasks (Table 9b as opposed to 9a).

Table 9A. Similarity: Object Visualization Ability, Low vs High- Graphs

	Task Type		
Object Vis Ability	Computational	Holistic	Domain-Specific
S (sig)	.30 (.237)	.39 **(.047)** *	.35 (.338)

Note. * Indicates significant similarity. † Indicates significant difference.

Table 9B. Similarity: Object Visualization Ability, Low vs High- Grables

	Task Type		
Object Vis Ability	Computational	Holistic	Domain-Specific
S (sig)	.40 **(.020)** *	.33 (.355)	.44 **(<.001)** *

Note. * Indicates significant similarity. † Indicates significant difference.

Table 10A. ΔCoherence Low-to-High: Object Visualization Ability- Graphs

	Task Type		
Object Vis Ability	Computational	Holistic	Domain-Specific
ΔC Low to High	-.71	.96	.79

Note. Coherence does not have a significance value.

Table 10B. ΔCoherence Low-to-High: Object Visualization Ability- Grables

	Task Type		
Object Vis Ability	Computational	Holistic	Domain-Specific
ΔC Low to High	.04	-.88	0

Note. Coherence does not have a significance value.

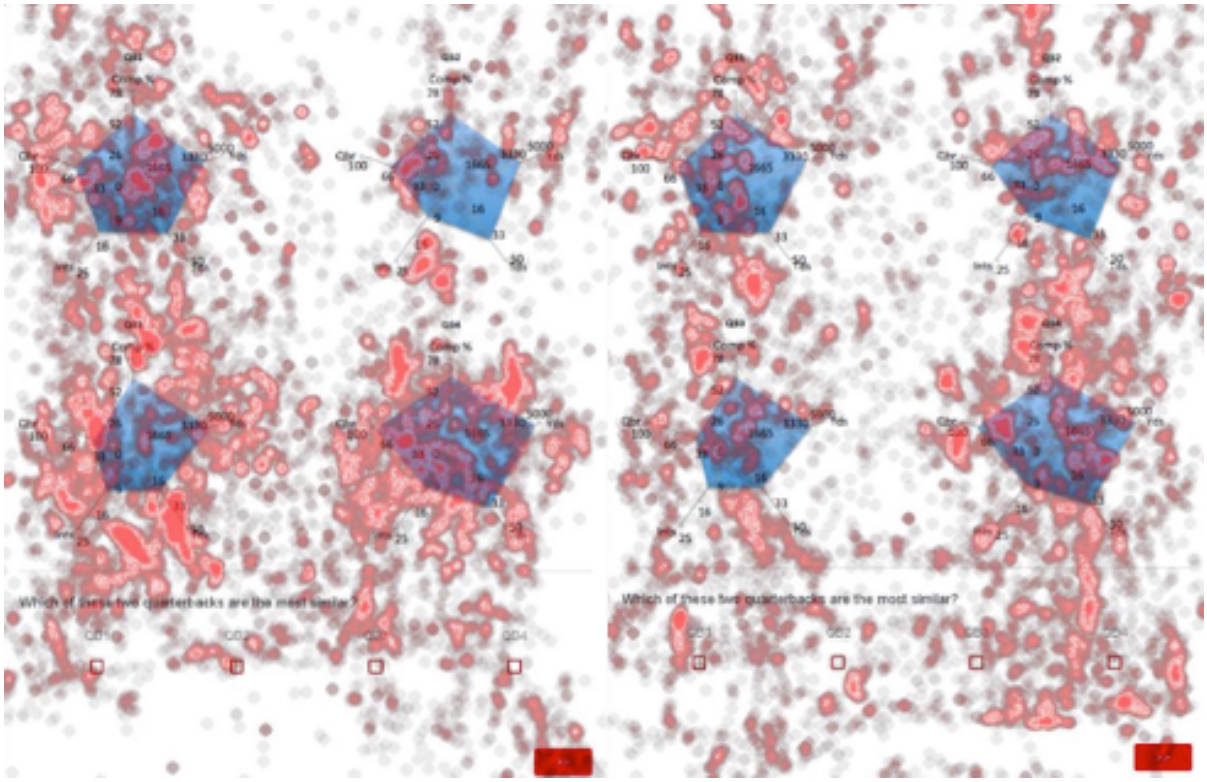


Figure 7a. Comparison of High (left) and Low (right) Object Visualization Ability On Holistic Tasks- Heatmaps

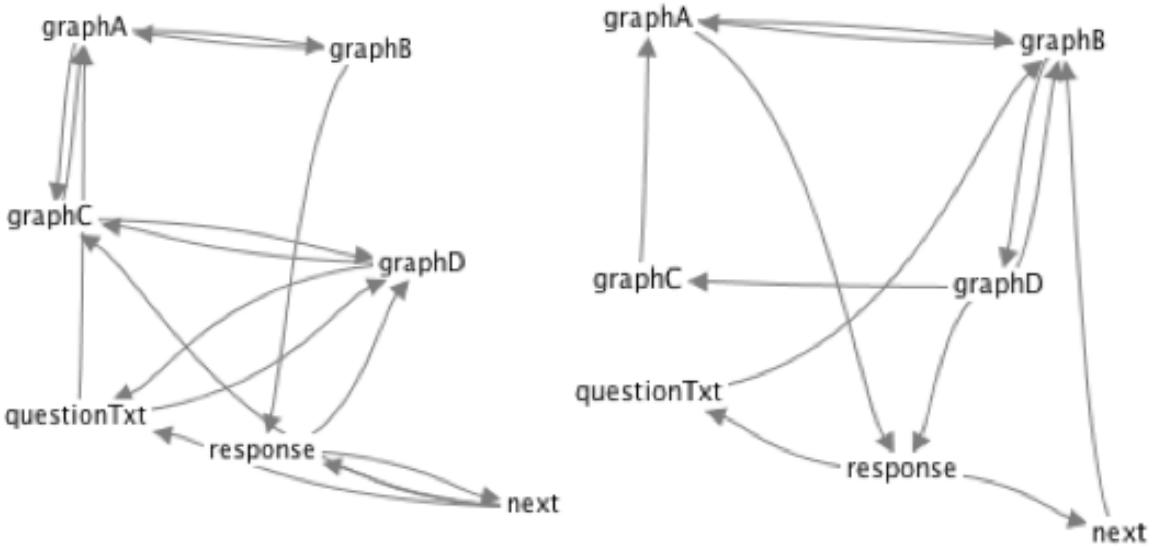


Figure 7b. Comparison of High (left) and Low (right) Object Visualization Ability On Holistic Tasks- Pathfinder Networks

Regression: Coherence Related to Response Time

An additional analysis was performed to observe if coherence had any relationship with response time for a given trial. Researchers were particularly curious about this relationship, as it would show a link between search pattern and performance on task. For this analysis, coherence (for each trial) was used as the sole predictor variable in a regression equation with response time as the dependent variable. Coherence was a significant predictor of response time, $b = .08$, $t(462) = 3.02$ and explained a significant, yet small proportion of the variance in response time ($R^2 = .019$, $F(1, 462) = 9.09$, $p = .003$).

Discussion- Hypothesis I

Taken together these results partially support hypothesis one. The study showed group-level differences in coherence for user variables. Differences in the coherence of a scan pattern, rather than the specific links in the scan pattern, seem to match up with the user-

related characteristics measured. For instance, the high expertise group's coherence in domain specific tasks is much higher than non-experts. Additionally, these results show that adding labels fundamentally changes the consistency (as function of task type whereby coherence was increased for computational tasks but decreased for holistic tasks) and pattern (become more similar) with which users scan through graphs. Finally, these results supported the proposal that differences in search can account for differences in performance (in terms of response time) for these graph-reading tasks.

Hypothesis II: Visualization Ability Will Interact With Task Type and Affect Performance and Search

To test hypothesis II, three different analyses were required; these are described in Table 11 below. In addition to this set of analyses, results from the PRONET analyses (from Hypothesis I) help to answer whether or not visualization ability has any effect on search. The results of PRONET (see Table 7a) analyses revealed that high and low spatial visualizers have a significantly different latent scan pattern in computational tasks, but not holistic or domain-specific tasks. On the other hand, there were no scan pattern differences found for different levels of object visualization ability (see Table 9a). Internal consistency of these scan pattern saw a similar pattern for both visualization abilities whereby those who score low had higher coherence on computational tasks, but participants who had high visualization abilities had higher coherence on holistic and domain-specific tasks (Tables 8a and 10a). So, despite finding scan pattern differences only in one setting (spatial visualization

ability * computational tasks) one can point to coherence (consistency) as a potentially different quality of scan pattern accounted for by visualization ability.

Table 11. Questions and Corresponding Analyses for Hypothesis II

Question	Analysis
What effect, if any, does visualization ability have on response times for the experimental tasks?	ANOVA IVs: object & spatial visualization ability DV: Response time (all tasks)
Is there a relationship between visualization ability and accuracy (for computational tasks)?	Logistic Regression IVs: object & spatial visualization ability DV: accuracy on computational task
Do people with different visualization abilities respond differently to holistic tasks?	Cramer's V IV: object & spatial visualization ability DV: response choice on Holistic task

Response Time by Visualization Ability and Task Type ANOVA

To test visualization ability's effects on reaction time, an ANOVA using response time as the dependent measure and object visualization group, spatial visualization group, and task as independent measures was used. The test results showed significant main effects of object visualization ability ($F(2, 837)= 8.37, p < .001, \eta^2= .02$), spatial visualization ability ($F(2, 837)= 11.03, p < .001, \eta^2= .026$), and task type ($F(2, 837)= 58.23, p < .001, \eta^2= .122$), indicating that all three of these variables have a significant effect on response times. Post-hoc tests were calculated to examine the differences. These post hoc tests for task show that the computational task ($M= 43.52$) took significantly longer than domain-specific tasks ($M= 24.31$) and holistic tasks ($M= 29.01$); additionally, the difference between the holistic and domain-specific task were significant. These task times are depicted in Figure 8 below.

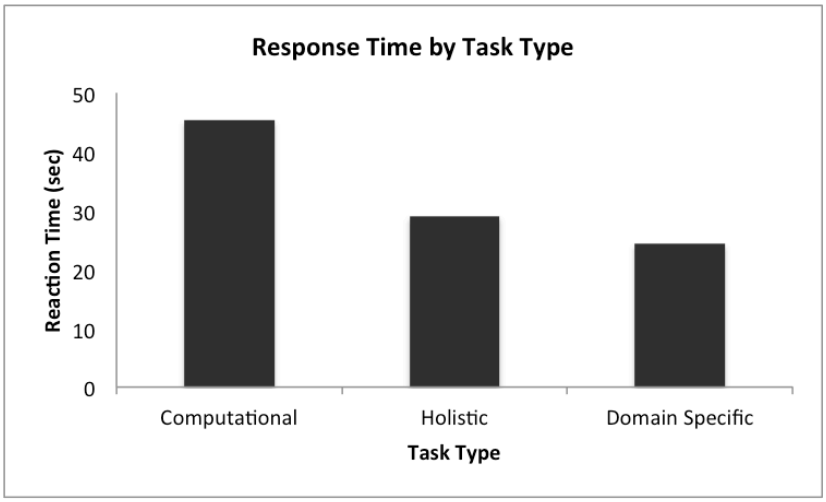


Figure 8. Mean response times by task type.

For spatial visualization ability, the low group (*M*= 29.01) and the high group (*M*= 30.20) were significantly faster than the middle group (*M*= 37.61), however there was no difference between the low and high groups meaning that those with high or low scores on spatial visualization ability perform tasks faster than those with middle-range scores. These data are shown visually in Figure 9 below.

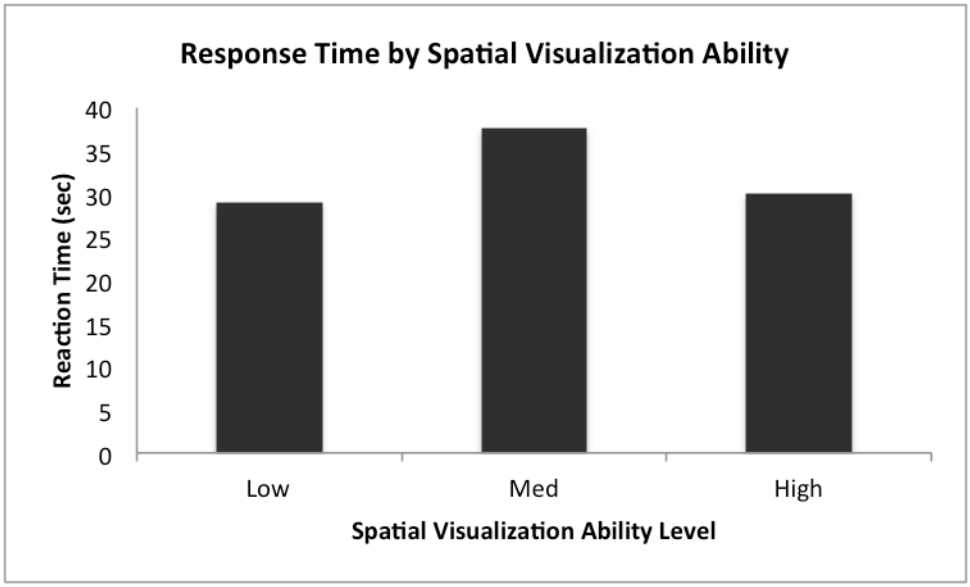


Figure 9. Mean response time by Spatial Visualization Ability Level.

For object visualization ability, the low (*M*= 28.47) and middle (*M*= 31.92) groups were significantly faster than the high group (*M*= 36.02), but not significantly different from each other; these numbers are shown below in Figure 10. In other words, people with higher object visualization ability took longer than all others for the tasks tested.

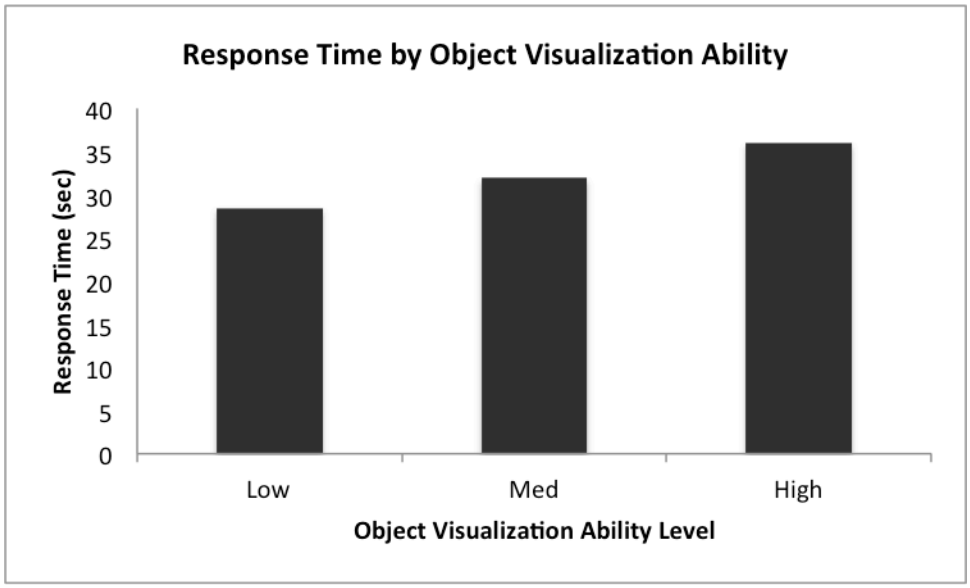


Figure 10. Response Time by Object Visualization Ability.

Interactions were also assessed to see if there were any effects that varied depending on the influence of another variable. No two-way interactions were significant (Object visualization ability * Spatial Visualization ability, $F(4, 837)= 1.73$, $p = .142$; Spatial visualization ability * Task, $F(4, 837)= 1.48$, $p = .207$; & Object visualization ability * Task, $F(4, 837)= 2.36$, $p = .052$). However, there was a significant three-way interaction between object visualization ability, spatial visualization ability, and task ($F(8, 837)= 2.351$, $p = .017$, $\eta^2= .022$). This three-way interaction is visualized in Figures 11a, 11b, and 11c where the x axis shows task type, the y-axis shows response time, separate lines show levels of spatial

visualization ability and different graphs show different levels of object visualization ability. The trends in response time for low and mid-range scores on spatial visualization ability stay fairly consistent between levels of object visualization ability (figure 11a, 11b). However, for those who scored high on spatial visualization ability, object visualization ability affects response time differently across different tasks. When a subject has high scores on both object and spatial visualization ability (compare Figure 11c to 11a and 11b), their time-to-respond for holistic and domain specific tasks were very low, but their response time for the computational task was very high. This finding suggests that people who have high scores on both visualization scales may have different ways of processing visual information or engage in different strategies in assessing visual information than other individuals. When a subject has high scores on spatial visualization ability and medium score on object visualization ability their response times follow a similar pattern to other participants. Finally, when a subject has low object visualization scores but high spatial visualization scores they are generally slightly faster for both computational tasks and holistic tasks however they are somewhat slowed by domain specific tasks.

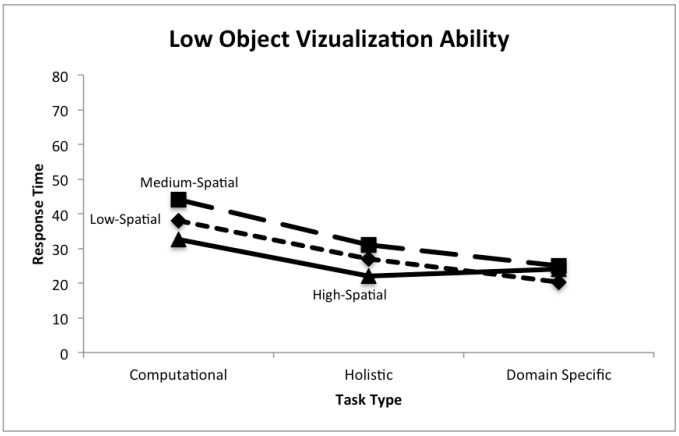


Figure 11a.

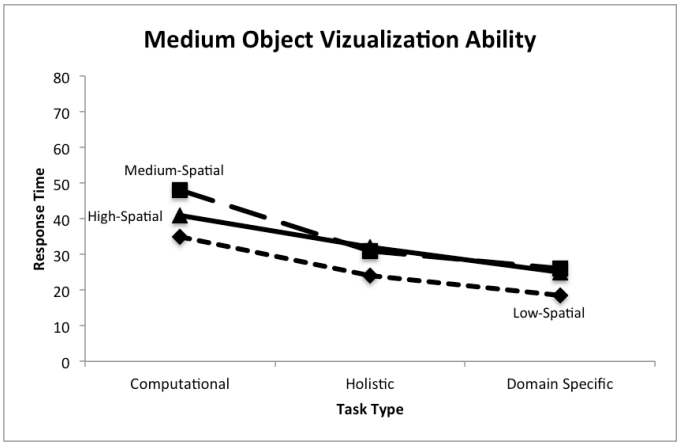


Figure 11b.

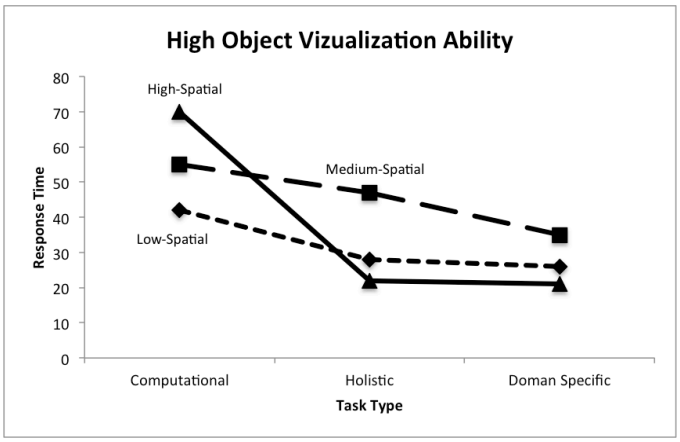


Figure 11c.

Accuracy by Visualization Ability Logistic Regression

Accuracy on the computational task is another measure in which visualization ability may affect performance. Accuracy for the computational task was analyzed using a logistic regression where responses within +/- 5% of the actual answer were recorded as correct (e.g., if the correct response was 25, responses between 23.75 and 26.25 were accepted). The overall model can be found in Appendix F (see Table 14). These analyses indicated that neither object visualization ability (Exp(β)= .88, Wald= .21, p= .647) or spatial visualization ability (Exp(β)= 1.08, Wald= .07, p= .791) were significant individual predictors of accuracy with computational tasks. In other words, neither object nor spatial visualization ability were found to make the graph reader more or less accurate in making computations using graphical aids (once other variables, like a graph-reader's mathematical abilities, are controlled for statistically).

Holistic task response by visualization ability Cramer's V

The responses on the holistic task (which of these two quarterbacks are the most similar?) were analyzed to assess whether visualization ability has an effect on response for tasks that can be processed holistically. A Cramer's V was used to analyze whether the choice on the holistic task differed by visualization ability level. Note that each trial was run as a separate test because the stimuli in each trial pairing were different. Based on this analysis the choice of the most similar quarterbacks was not influenced by either spatial visualization ability (Φ= .27, p= .336; Φ= .28, p= .505; Φ= .34, p= .38) or object visualization ability (Φ= .31, p= .158; Φ= .28, p= .511; Φ= .31, p= .541). In other words,

this study found no evidence that the response to a holistic task depends on a person's visualization abilities.

Discussion- Hypothesis Two

These findings taken together partially support hypothesis two.

- Supported—time on task was shown to be influenced by an interaction of the two types of visualization ability and task *(as evidenced by 3-way ANOVA interaction)*
- Failed to support—accuracy on computational tasks was not influenced by either object or spatial visualization ability *(as evidenced by logistic regression)*
- Failed to support – answer choices on holistic tasks were not influenced by either object or spatial visualization ability *(as evidenced by Cramer's V)*
- Partially supported—visual scan pattern was different between high and low spatial visualizers only for computational tasks; no other significant differences in scan pattern were found. Additionally, high scores on visualization ability (either object or spatial) are associated with lower coherence for computational tasks, but higher coherence for holistic and domain-specific tasks *(as evidenced by PRONET analyses)*

These analyses have important implications for our understanding of graph reading and the overall visualization ability literature. Response time was shown to be a three-way interaction of task type, spatial visualization ability and object visualization ability, indicating that both forms of visualization ability are important and have interactive effects that may manifest themselves differently under different task conditions. Consequently,

researchers studying visualization ability and mental imagery should measure both sub-abilities to determine their joint effects.

Analysis also revealed that accuracy on computational tasks and responses on holistic tasks were not affected by visualization ability. Additionally, although visualization ability is shown to be an important factor for response times, simply training to improve a user's visualization abilities would not guarantee that their general graphical performance would improve. Rather, visualization training would need to be matched with task demands (e.g. time constraints, task type, response modality).

Finally, through eye tracking it was observed that (other than spatial visualization ability and computational tasks) visualization ability does not impact how people scan through graphical aids. However, there appear to be important differences in coherence whereby high visualization abilities lend themselves to higher consistency in holistic and domain specific tasks, while low visualization abilities tend to have higher coherence in computational tasks. Future research on coherence is needed for both validation of these results and to explore the full relationship between it and user, task, and performance variables.

Hypothesis III: Expertise Will Affect Performance and Search in Domain Specific Tasks

Testing Hypothesis III required a set of three analyses. These analyses and the questions they are intended to answer are described in Table 12 below. In addition to this set of analyses, results from the PRONET analyses (found in Hypothesis I) help to answer

whether or not expertise has an effect on visual search. The results of PRONET (see Tables 3a, 5a) showed that low and high expertise individuals had statistically *similar* scan pattern for domain specific tasks with graphs. Despite their similarity in overall scan pattern the high expertise group had higher coherence (seen in tables 4a and 6a) than did the low-expertise group (for survey-based expertise high-expertise C=.77, low-expertise C= -.07; for reported hours based expertise high C= .76, low C= -.48). This higher coherence coupled with a statistically similar scanning pattern indicates that people with different levels of expertise with a subject matter may scan through graphs similarly, however those with high expertise will have a greater overall consistency.

Table 12. Questions and Corresponding Analyses for Hypothesis III

Question	Analysis
Does expertise have an effect on response time for graphical tasks where the data displayed pertains to their domain-expertise?	ANOVA IV: expertise (measured by survey) and expertise (measured by reported hours spent) DV: response time (all tasks)
Does expertise influence accuracy in computational tasks where the information displayed regards domain knowledge?	Logistic Regression IV: expertise (measured by survey) and expertise (measured by reported hours spent) DV: accuracy (computational task)
Does expertise influence response choice in domain specific tasks?	Cramer's V IV: expertise (measured by survey) and expertise (measured by reported hours spent) DV: response on domain-specific task

Response Time by Expertise and Task ANOVA

To analyze the effect of expertise on response times, an ANOVA using both expertise measures – survey responses and reported weekly hours engaging in football-related activities– was conducted. Expertise measured by survey item responses ($F(2, 843)= 10.89, p < .001, \eta^2= .025$) and weekly reported hours ($F(2, 843)= 9.21, p < .001, \eta^2= .021$) have a

significant effect on response times. This finding suggests that expertise significantly affects response times for tasks involving graphical aids. However, post-hoc tests revealed a conflicting pattern—for reported weekly hours, the medium group completed tasks significantly more quickly than the low and high scoring groups. In contrast, for expertise as measured by survey items, the high group completed tasks significantly more quickly than low and medium groups. This difference in results as a function of how expertise was measured could be due to reporting error (for weekly hours). Alternatively, the difference could be due to those who spent more time in weekly football related activities also spent more time trying to decompose the test tasks. There were no significant interactions between task type and expertise, suggesting that domain expertise may benefit response times on all tasks rather than only on tasks that specifically require domain expertise.

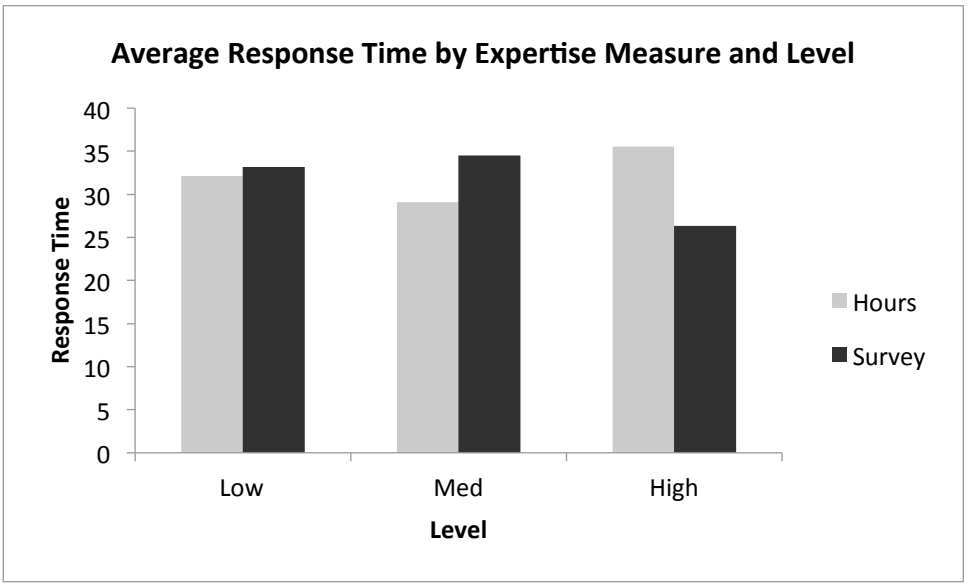


Figure 12. Response Time by Expertise Measure and Level.

Accuracy by Expertise Logistic Regression

Accuracy (for the computational task) was analyzed using a logistic regression in which responses within +/- 5% of the actual answer were recorded as correct. The overall model can be found Appendix F (Table 14). In this model, neither expertise as measured by survey items (Exp(β)= 1.01, Wald= .27, *p*= .604) nor as measured by hours-per-week-on-tasks (Exp(β)= 1.00, Wald= .04, *p*= .837) were found to be significant individual predictors of accuracy with computational tasks. Thus, no evidence was found to suggest that domain-expertise has an effect on a person's accuracy in performing computational problems using graphical aids.

Response on domain specific task by expertise Cramer's V

Finally, an analysis examined whether response on the domain specific task was affected by expertise level. A Cramer's V was used to analyze whether the answer choice selected was different by expertise level for each domain-specific trial. As in the analysis of visualization ability, each trial was run as a separate test because the stimuli in each trial differed. Based on this analysis, the answer choice selected was not influenced by expertise whether it is measured by survey responses (Φ= .25, *p*= .185; Φ= .21, *p*= .398; Φ= .20, *p*= .414) or hours of experience (Φ= .25, *p*= .221; Φ= .18, *p*= .523; Φ= .23, *p*= .269). Results show no evidence that domain expertise has an impact on the given response for a domain-specific task using graphical aids. This finding is surprising, given that experts should be able to apply their domain-knowledge to these problems and come up with different responses than their novice counterparts. Perhaps the expertise manipulation was too weak or the

football domain was so familiar that even the novice participants could determine how to respond to the tasks.

Discussion Hypothesis Three

Taken together these results partially support hypothesis three.

- Partially Supported – time on task was shown to be affected by expertise level, however the results were inconsistent. When survey items measure expertise, experts were fastest. However, when the measure for expertise was self-reported time spent per week, the medium group performed tasks faster than high or low expertise groups. *(as evidenced by ANOVA)*
- Failed to support – expertise did not have a significant impact on accuracy *(as evidenced by logistic regression)*
- Failed to support – expertise level did not have a significant effect on response choices for domain specific tasks *(as evidenced by Cramer's V)*
- Partially supported – expertise did not have a significant impact on scan pattern, in that experts and non-experts had statistically similar scan patterns for domain-specific tasks. However, the consistency (coherence) of expert's scan patterns were higher than non-experts for all tasks, and especially in domain-specific tasks

Time to complete graph reading tasks was affected by level of expertise and a broad effect was found for scan pattern consistency. However, expertise had no affect on accuracy in computations (much like visualization ability), or response choice on domain specific tasks. These findings suggest that any evaluation techniques for information visualization, and

models of graph reading need to account for domain expertise, but not as widely as hypothesized. These results have important implications for training as well; developing domain expertise would have a positive impact on performance in tasks using graphical aids and benefit in scanning reliability. This result was particularly interesting because it suggests that expertise in a domain area has a benefit for any task, not just questions where expertise can be applied (like computations for example) involving a graphical aid (potentially even an unfamiliar one).

Hypothesis IV: Performance Will Be Improved by Grables

Testing hypothesis IV required a set of four analyses, as summarized below in Table 13. It is also worth noting that paired Coherence values for a given group (i.e. taking the mean group coherence values for a group with graphs and the same group with grables) had significant results for computational tasks (U= 8, z=-2.47, p= .013) and holistic tasks (U=10.5, z= 2.21, p= .027). However, the effects were in opposite directions. Using grables in computational tasks increased coherence in the Pathfinder analyses of scan patterns over use of graphs. Thus, for computational tasks, graphical labels improved scanning consistency. In contrast, for holistic tasks, the addition of graphical labels had a negative impact on the coherence of most groups. There was no effect of graph type on domain specific tasks (U= 27, z= .47, p= .638). Additionally when comparing similarity values between display types one can see that there is a fairly global effect whereby grables yield scanning patterns with higher similarity than graphs (Z= 2.31, p= .021).

Table 13. Questions and Corresponding Analyses for Hypothesis IV

Question	Analyses
Do grables improve performance in terms of time on tasks?	ANOVA IV: display type; user variables, task type DV: response time
Do grables improve performance in terms of accuracy on computational tasks?	Logistic Regression IV: display type DV: accuracy
Do grables have an effect on responses on holistic tasks?	Cramer's V IV: display type DV: response choice on holistic task
Do grables have an effect on responses to domain-specific tasks?	Cramer's V IV: display type DV: response choice on domain-specific task

Response time by display and task ANOVA

Display type was not a significant predictor of response times ($F(1, 858)= 1.05$, $p= .306$, $\eta^2= .001$). Further, no significant interactions with display types were found with task ($F(2, 858)= .10$, $p= .904$, $\eta^2= .000$), expertise as measured by reported hours per week ($F(2, 822)= 1.66$, $p= .190$, $\eta^2= .004$), expertise as measured by survey items ($F(2, 822)= 1.87$, $p= .155$, $\eta^2= .005$), object visualization ability ($F(2, 810)= .36$, $p= .710$, $\eta^2= .001$), or spatial visualization ability ($F(2, 810)= .72$, $p= .489$, $\eta^2= .002$). Taken together these results show that grables show no benefit over graphs in terms of response times for any of the tasks measured, nor do they benefit any particular group (high visualizers, experts, etc.).

Accuracy by Display Logistic Regression

Display type's affect on accuracy (for the computational task) was analyzed using a logistic regression where responses within +/- 5% of the actual answer were recorded as correct. The overall model can be found in a previous section (see Table 3). In that model, display type ($Exp(\beta)= 9.67$, Wald= 57.04, $p< .001$) was found to be a significant predictor of

accuracy with computational tasks, indicating that participants were far more accurate with grables than with graphs (see Figure 13).

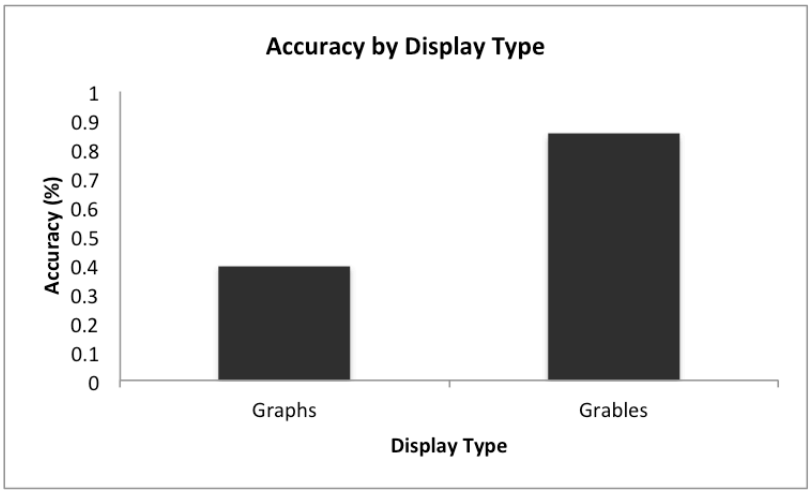


Figure 13. Accuracy by Display Type.

This result is not surprising given that grables eliminate the step of having to accurately interpret and assign values to graphical indicators. Instead the participants could simply search for the needed values and make the exact mental calculations. With graphs participants were far less accurate, but took almost equal time to respond to computational tasks, suggesting that they may have focused more of their efforts on assigning values to the graph and then engaging in rough math. This effect is important as it may indicate that graphs prime a viewer towards general, rough-estimates of calculations, while grables prime arithmetic.

Response For Domain-Specific Task by Display Cramer's V

Next, responses to the domain-specific task were analyzed to see if domain expertise had a significant effect on response choice. This analysis was achieved using a Cramer's V. There were 6 domain-specific task trials in the experiment, consisting of three pairs of

stimuli that were identical visuals with the exception of graphical labels, making for three sets of comparisons. Based on this analysis, the answer choice was influenced by display type for two of three trial pairings ($\Phi = .07, p = .971$; $\Phi = .491, p = .001$; $\Phi = .528, p < .001$). This was an inconsistent effect and needs more future investigation to validate; however, it suggests that there *may* be an effect on decision-making whereby providing exact values in graphical aids changes how a graph reader responds in a domain-specific task.

Response On Holistic Task by Display Cramer's V

Finally, researchers tested whether display type would affect responses on the holistic task, which might indicate that providing exact values on a graph would change how readers choose to either engage in holistic or analytical thinking. Responses were analyzed using a Cramer's V using answer choice as the outcome variable and display type as the independent variable. There were 6 holistic task trials in the experiment, consisting of three pairs of stimuli that were identical graphs, making for three sets of comparisons. Based on this analysis the choice of the most similar pair was not influenced by display type for any of the trail pairings ($\Phi = .30, p = .211$; $\Phi = .33, p = .284$; $\Phi = .33, p = .466$), indicating that there is no evidence that grables (in comparison to regular graphs) influence a reader to engage in different holistic or analytical processes.

Discussion- Hypothesis IV

Taken together these results partially support hypothesis IV.

- Failed to support – display type was not shown to be a significant predictor of response times *(as evidenced by ANOVA)*

- Supported – display type was shown to be a significant predictor for accuracy, whereby trails where grables were provided yielded greater accuracy *(as evidenced by logistic regression)*
- Partially supported – display type appears to have some effects on answer choices for the domain specific task *(as evidenced by Cramer's V)*
- Failed to support – display type had no significant effect on the holistic task *(as evidenced by Cramer's V)*

  Partially supported – display type had an inconsistent effect on coherence that depended on task and user-characteristic situational variables. However, there was a general trend whereby grables (as opposed to graphs) yielded statistically similar scan patterns between high and low scoring groups for both visualization ability and expertise for all tasks *(as evidenced by PRONET analyses)*

The results from hypothesis IV have clear implications for guidelines for designing and assessing the usefulness of a graphical aid for a given task. If the user's task values accurate computations, grables should be provided instead of graphs. However, if a task involves discovery or a team of people who need to make several different observations graphs, should be provided instead of grables to promote a more varied scanning pattern.

GENERAL DISCUSSION

The results of the present experiments at least partially supported all of the hypotheses proposed. Starting with data from eye tracking, the experiments showed that there

are potential differences in scanning based on user, task and display related factors. However, the difference (in most cases) was not in latent scan pattern as originally hypothesized. Rather, it was the coherence – or internal reliability—of the transitions from fixation-to-fixation where differences were noted. Additionally, higher coherence was linked to lower time on task, establishing a link between scanning consistency and performance outcomes. Exploring further relationships between coherence and other performance outcomes in a variety of different tasks is a potential opportunity for future work.

The eye tracking results, which showed that, largely, the actual scanning pattern did not differ between the various conditions, lends support to Lohse's (1991 & 1993) semantic trace theory which proposed that there is one logical scan pattern that is more-or-less followed for a given task. One basis for the present research was to test the idea that user variables would account for scan pattern differences. However the results tended to support a single scan pattern approach. The one caveat that this research adds is that, although the same scan pattern is generally followed differences in user-related, and display-type variables will account for differences in the reliability of the transitions from fixation-to-fixation in the scan paths (as seen by the coherence data collected).

Another important takeaway from this experiment is that all the factors that affect graph reading are extremely task dependent. User-level differences only show significant effects under the right circumstances (i.e. task and display conditions). For example, domain expertise increased coherence (or scanning consistency) only in domain-specific tasks, and spatial visualization ability increased coherence only in holistic tasks. This task and user

specificity has important implications for any efforts towards developing a set of analysis and evaluation techniques for graphs—use case would become an important factor, as it would determine which user related characteristics have potential to impact performance and usability. This would also mean that any training of skills for graph reading would have to be situation dependent as well, as training one skill may have a benefit in one task setting but have a negative impact on a different task.

As for the user-related characteristics that most impact performance, general mathematical ability was the most important factor for tasks that involve making calculations. Domain expertise benefits the user in graph reading tasks, especially on domain-specific tasks, but also on other tasks where their domain knowledge is displayed. This experiment also establishes that both types of visualization ability (object and spatial) are important considerations for graph reading tasks, as they individually affect outcomes, but also interact with each other in certain situations. This has further implications for basic research on visualization ability, where, more-often-than-not, the focus is only spatial visualization ability (i.e. scanning, rotation). More nuanced effects might be found if both subscales of visualization ability were measured. For example, in the present research, those who scored highly on both subscales have appreciably different performance from other participants.

The present experiment also showed support for providing exact numerical labels on graphs (i.e. grables) in that they make computations more accurate. However, the results also show that these exact numbers impact the scan patterns for some groups—and makes for a

fairly homogenous style of scanning. Consequently, visualization makers should be careful in considering whether or not to provide exact labels. If a task requires any computations, grables should be used as graphical aids, but for other tasks they may not be valuable, or perhaps even detrimental to some users.

The experiment reported here establishes visualization ability, and expertise as user-related characteristics of interest, however there are likely other measures (e.g. aspects of attention or cognition) that could impact graph-reading. Likewise, a multitude of display factors other than graphical labels, and a number of other task types that will impact human-graph interaction. So the field of human-graph interaction has many more research questions to address. For example, follow-on research from the present study might focus on the manipulations performed here, but when the user is under time pressure.

Further research should strive to replicate the effects found in regards to the homogeneity of scan pattern and coherence as the primary indicator of scanning differences between graph readers of varying ability levels of a given attribute. Additionally, further research should investigate how medium of delivery (i.e. computer screen, large-screen, small-screen, print-out, embedded images, etc.) may impact graph-reader interpretations, as these different display sizes impact the perceived size of the graphical indicators.

REFERENCES

Amar, R., Eagan, J., & Stasko, J. (2005). Low-level components of analytic activity in information visualization. *Proceedings of IEEE InfoVis*, 111-117.

Blajenkova, O., Kozhevnikov, M., & Motes, M. A. (2006). Object-Spatial imagery: A new self-report imagery questionnaire. *Applied Cognitive Psychology,* 20, 239-263.

Borwein, J. & Bailey, D.H. (2015). Moore's Law is 50 years old but will it continue? Phys Org; Retrieved from: http://phys.org/news/2015-07-law-years.html

Chase, W.G., & Simon, H.A. (1973). Perception in chess. *Cognitive Psychology, 4,* 55-81.

Dipert, B. (2010). Display-technology advancements: Change is the only constant. EDN Network. Retrieved from: http://www.edn.com/design/power-management/4363883/Display-technology-advancements-Change-is-the-only-constant

Ekstrom, R. B., French, J. W., & Harman, H. H. (1976). Kit of factor-referenced cognitive tests. Princeton, NJ: Educational Testing Service.

Freedman, E.G., Shah, P.S. (2001). Individual differences in domain knowledge, graph reading skills, and explanatory skills during graph comprehension. *Paper Presented at 42nd Annual Meeting of the Psychonomic Society,* Orlando, FL.

Gattis, M., & Holyoak, K.J. (1996). The role of theory and data in covariation assessment: Implications for the theory-ladennes of observation. *Journal of Mind and Behavior*, *17*, 321-343.

Goldberg, J.H., & Helfman, J.I. (2010). Comparing information graphics: a critical look at eye tracking. In *Proceedings fo the 3rd BELIV'10 Workshop: Beyond time and errors: Novel evaluation methods for information visualization* (pp. 71-78). ACM

Gillan, D.J. (2000). A componential model of human interaction with graphs: V. Using pie graphs to make comparisons. In *Proceedings of the IEA 2000/HFES 2000 Congress* (pp 3-439 - 3-443). Santa Monica, CA: HFES.

Gillan, D.J. (2009). A componential model of human interaction with graphs. VII: A review of the mixed arithmetic-perceptual model. *Proceedings of the Human Factors and Ergonomics Society annual meeting* (Vol. 53, No. 12, pp.829-833.) SAGE Publications

Gillan, D. J. & Calahan, A. B. (2000). A componential model of human interaction with graphs.VI: cognitive engineering of pie graphs. *Human Factors,* 42, 566-591.

Gillan, D.J., & Cooke, N.J. (2001). Using pathfinder networks to analyze procedural knowledge in interaction with advanced technology. *Advances in Human Performance and Cognitive Engineering Research, 1,* 125-162.

Gillan, D.J. & Harrison, C. (1999). A componential model of human interaction with graphs. IV: Holistic and analytical perception of star graphs. *Proceedings of Human Factors and Ergonomics Society annual meeting* (Vol. 43, No. 23, pp. 1304-1307). SAGE Publications.

Gillan, D.J. & LaSalle, M.S. (1994). A componential model of human interaction with graphs. III. spatial orientation. *Proceedings of Human Factors and Ergonomics Society annual meeting* (Vol. 38, No. 4, pp. 285-289). SAGE Publications.

Gillan, D.J. & Lewis, R. (1994). A componential model of human interaction with graphs: 1. Linear regression modeling. *Human Factors 36*(3), 419-440.

Gillan, D.J., & Neary, M. (1992, October). A Componential Model of Human Interaction with Graphs. H. Effects of the Distances among Graphical Elements. *In Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 36, No. 4, pp. 365-368). SAGE Publications.

Glaser, R. (1987). Thoughts on expertise. In C. Schooler & W. Schaie (Eds.), *Cognitive functioning and social structure over the life course*. Norwood, NJ: Ablex.

Goldberg, J., & Helfman, J. (2011). Eye tracking for visualization evaluation: Reading values on linear versus radial graphs. *Information Visualization, 0*(0), 1-14.

Hink, J.K., Wogalter, M.S., Eustace, J.K. (1996). Display of quantitative information: are grables better than plain graphs or tables? *In Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 40, No. 23, pp. 1155-1159). SAGE Publications

Hirsch, S.G. (1997). How do children find information on different types of tasks?: Children's use of the science library catalog. *Library Trends, 45*(4), 725-745.

Huang, W. (2007). Using eye tracking to investigate graph layout effects. In Visualization, 2007. *APVIS'07 6th International Asia-Pacific Symposium*. (97-100). IEEE

Imagery Testing Battery (Version 1.0) [Computer Software]. Newark, NJ: MM Virtual Design, LLC: www:mmvirtualdesign:com

Interlink Inc. (2017). JPathfinder.

Kosslyn, S. M. (1975). Information representation in visual images. *Cognitive Psychology,* 7, 341-370.

Kosslyn, S. M., Ball, T. M., & Reiser, B. J. (1978). Visual images preserve metric spatial information: Evidence from studies of image scanning. *Journal of Experimental Psychology: Human Perception and Performance,* 4, 47-60.

Kuhlthau, C. (1999). The role of experience in the information search process of an early career information worker: Perceptions of uncertainty, complexity, construction and sources. *Journal of the American Society for Information Science, 50*(5), 399-412.

Liu, S., Cui, W., Wu, Y., & Liu, M. (2014). A survey of information visualization: recent advancements and challenges. *The Visual Computer, 30*(12), 1373-1393.

Lohse, G.L. (1991). A cognitive model for the perception and understanding of graphs. *Proceedings of the SIGCHI conference on Human factors in computing systems* (137-144).

Lohse, G.L. (1993). Eye movement-based analyses of graphs and tables: the next generation. *ICIS* 213-224.

Marchionini, G. (1995). Information seeking in electronic environments. Cambridge, UK: Cambridge University Press.

Marchionini, G., Dwiggins, S. Katz, A., & Lin, X. (1993). Information seeking in full-text and end-user oriented search systems: The roles of domain and search expertise. *Library and Information Science Research, 15*(1), 35-69.

Marks, D. F. (1973). Visual imagery differences in the recall of pictures. *British Journal of Psychology*, 64, 17–24.

McDonald, S., & Stevenson, R.J. (1998). Navigation in hyperspace: An evaluation of the effects of navigation tools and subject matter expertise on browsing and information retrieval in hypertext. *Interacting with Computers, 10,* 129-142.

Meyer, J., Shinar, D., & Leiser, D. (1997). Multiple factors that determine performance with tables and graphs. *Human Factors,* 39(2), 268-286.

Patel, S.C., Drury, C.G., & Shalin, V.L. (1998). Effectiveness of expert semantic knowledge as a navigational aid within hypertext. *Behavior and Information Technology, 17*(6), 313-324.

Peebles, D., & Cheng, P.C. (2003). Modeling the effect of task and graphical representation on response latency in a graph reading task. *Human Factors, 45*(1), 28-46.

Raven, J., Raven, J. C., & Court, J. H. (1998). Manual for Raven's advanced progressive matrices. *Oxford: Oxford Psychologists Press*.

Schvaneveldt, R.W. (1990). Pathfinder associative networks: Studies in knowledge organization. Ablex Publishing.

Shepard, R. N. & Metzler, J. (1971). Mental rotation of three-dimensional objects. *American Association for the Advancement of Science,* 171, 701-703.

Thomas, J., & Cook, K. (2005). Illuminating the path: the R&D agenda for visual analytics. National Visualization and Analytics Center, Institute of Electrical and Electronics Engineers.

Thomas, J.J. & Cook, K.A. (2006). A visual analytics agenda. *IEEE Computer Graphics and Applications,* 26, 10-13.

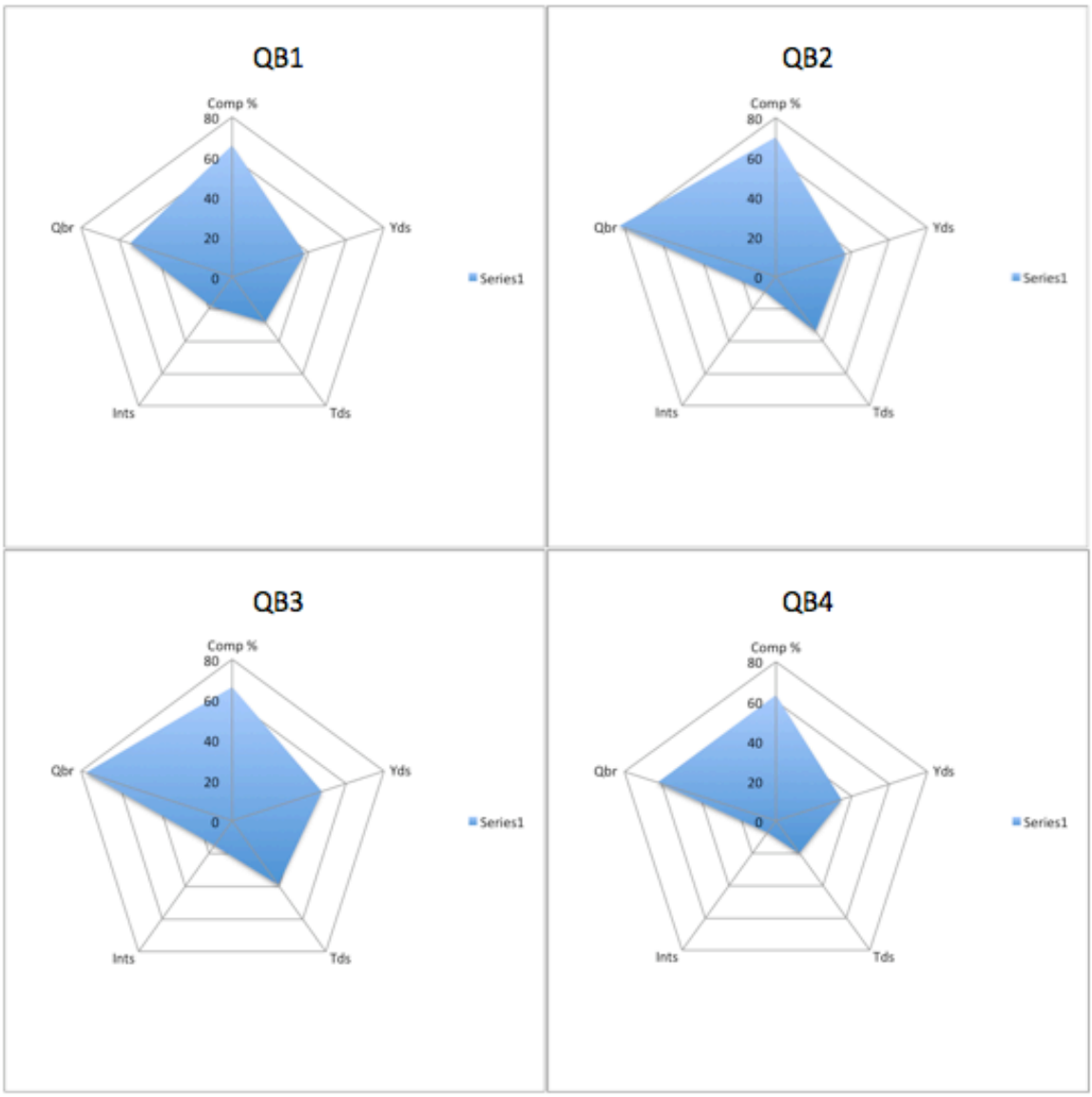
Qualtrics. (2017). Provo, UT, USA. Retrieved from http://www.qualtrics.com.

Vandenberg, S. G., & Kuse, A. R. (1978). Mental rotations, a group test of three-dimensional spatial visualization. *Perceptual & Motor Skills*, 47, 599–604.

APPENDICES

Appendix A

Example Trial Stimuli



Appendix B

Football Interest and Expertise Questionnaire

Items (Likert 1, strongly disagree to 5, strongly agree):

1. I watch football regularly

2. I am familiar with statistics associated with football and quarterback performance

3. I consider myself a football fan

4. My friends and family have interest in football

5. I enjoy conversing about football

6. I understand the difference between QBR and traditional passer rating

7. I play fantasy football regularly

8. I have experience - coaching or playing - football

Appendix C

OSIQ

Items (Likert 1, strongly disagree to 5, strongly agree):

1 I was very good in 3-D geometry as a student.

2 If I were asked to choose between engineering professions and visual arts, I would prefer engineering.

3 Architecture interests me more than painting.

4 My images are very colourful and bright.

5 I prefer schematic diagrams and sketches when reading a textbook instead of colourful and pictorial illustrations.

6 My images are more like schematic representations of things and events rather than detailed pictures.

7 When reading fiction, I usually form a clear and detailed mental picture of a scene or room that has been described.

8 I have a photographic memory.

9 I can easily imagine and mentally rotate 3-dimensional geometric figures.

10 When entering a familiar store to get a specific item, I can easily picture the exact location of the target item, the shelf it stands on, how it is arranged and the surrounding articles.

11 I normally do not experience many spontaneous vivid images; I use my mental imagery mostly when attempting to solve some problems like the ones in mathematics.

12 My images are very vivid and photographic.

13 I can easily sketch a blueprint for a building that I am familiar with.

14 I am a good Tetris player.

15 If I were asked to choose between studying architecture and visual arts, I would choose visual arts.

16 My mental images of different objects very much resemble the size, shape and colour of actual

objects that I have seen.

17 When I imagine the face of a friend, I have a perfectly clear and bright image.

18 I have excellent abilities in technical graphics.

19 I can easily remember a great deal of visual details that someone else might never notice. For example, I would just automatically take some things in, like what colour is a shirt someone wears orwhat colour are his/her shoes.

20 In high school, I had less difficulty with geometry than with art.

21 I enjoy pictures with bright colours and unusual shapes like the ones in modern art.

22 Sometimes my images are so vivid and persistent that it is difficult to ignore them.

23 When thinking about an abstract concept (e.g. 'a building') I imagine an abstract schematic building in my mind or its blueprint rather than a specific concrete building.

24 My images are more schematic than colourful and pictorial.

25 I can close my eyes and easily picture a scene that I have experienced.

26 I remember everything visually. I can recount what people wore to a dinner and I can talk about the way they sat and the way they looked probably in more detail than I could discuss what they said.

27 I find it difficult to imagine how a 3-dimensional geometric figure would exactly look like when rotated.

28 My visual images are in my head all the time. They are just right there.

29 My graphic abilities would make a career in architecture relatively easy for me.

30 When I hear a radio announcer or a DJ I've never actually seen, I usually find myself picturing what he or she might look like.

Appendix D

Further Information on OSIQ

In 2006 Blajenkova, Kozhevnikov, and Motes developed the OSIQ for assessment of both object and spatial factors of visualization ability. Many attempts were made to establish the reliability and validity of this measure. In the initial development of the scale the internal reliability (Cronbach's Alpha) was .83 and .79 for the spatial and object subscales respectively; the test-retest reliability for those subscales was r=.813 and r=.952 (both significant at p<.001 level). In a follow on study (reported in the same 2006 paper) Blajenkova, Kozhevnikov and Motes aimed to show that their subscales significantly related to accepted tests for assessing mental imagery and spatial abilities; their results yielded significant positive correlations between the spatial scale and PFT (Eckstrom et al., 1976), VK-MRT (Vandenberg & Kuse, 1978) and the spatial imagery test (Imagery Testing Battery Version 1.0), and between the object scale and the degraded pictures test and the VVIQ ((Marks, 1973). In further testing the researchers used the Advanced Progressive Matrices Test (Raven et al., 1998) to show that their scale was not significantly related to intelligence, displaying its construct validity. Additionally the researchers report findings regarding ecological validity, where they justified their measure by through administering it to professionals and then categorizing their careers by the skills they involve to be either higher in object or spatial visualization ability. They found that, typically, professionals that work in careers involving spatial imagery score high on that subscale while those who work mostly in object imagery score higher in that scale.

Appendix E

Arithmetic Assessment Instrument Items

	A	B	C	D	E
1 17 + 47 = 7 + ?	55	57	65	67	35
2 44 – ? = 15	26	29	28	39	30
3 7 × 8 = ?	49	56	64	54	52
4 140 ÷ 35 = ?	3	3.5	4	4.5	5
5 22% of 200 = ?	42	44	40	88	46
6 56.9 – 7.4 = ?	48.3	47.9	45.9	49.3	49.5
7 12.8 × ? = 3.2	0.20	0.25	0.30	0.33	0.40
8 A restaurant bill is made up of the following: $12.50 for starters, $28.55 for main courses and $8.95 for deserts, plus a 15% service charge. How much is the bill?	$56.50	$57.50	$57.00	$59.50	$60.50
9 A team of eight lumberjacks cut an average of 15,000 cubic feet of timber in a week. How many cubic feet will four lumberjacks cut in four weeks?	30,000	25,000	32,000	16,000	28,000
10 A discount of 15% is offered on an item which previously cost $1.80. What is the discounted price?	$1.53	$1.40	$1.55	$1.60	$1.52

Appendix F

Analyses on the Relationship Between Arithmetic Ability and Spatial Visualization Ability

As seen in Table 2, scores on the arithmetic assessment were significantly correlated with spatial visualization ability ($r = .459$, $p < .01$). This relationship could be problematic if all the significant effects due to spatial visualization differences are actually due to mathematical ability differences. Regression analyses were run to investigate this potential issue.

The first analysis involved a two-step regression with response time as the dependent variable using a two-step entry method, with arithmetic score added in the first step and spatial visualization score added in the second step. If spatial visualization ability adds nothing beyond being a proxy measure for arithmetic ability, then the R^2 difference (ΔR^2) for the two models would not be significant, or, alternatively, a significant ΔR^2 between the two models would indicate that spatial visualization ability adds a significant amount of descriptive power beyond mathematical ability alone. This regression analysis was performed twice; once for all trials and a second time including only trials on the computational task. The results of the regression analysis show that spatial visualization ability explains significantly more variance than a model that only includes mathematical ability for all tasks ($\Delta R^2 = .009$, $F = 7.52$, $p = .006$) and in computational tasks $\Delta R^2 = .025$, $F = 7.12$, $p = .008$), meaning that spatial visualization ability accounts for significantly additional variance in response time than mathematical ability alone and, thus, is a value-added predictor.

A similar analysis was done with accuracy data using logistic regression. For the purposes of this analysis any answer within +/- 5% of the actual calculated value was counted as correct. In the first step of the model mathematical ability was added; overall this step was significant, $\chi^2(1)=5.69$, $R^2= .03$, $p= .017$; and mathematical ability was a significant individual predictor of accuracy whereby higher mathematical abilities was associated with higher accuracy, Exp(β)= 1.21, Wald= 5.64, $p= .018$. In the second step, of the model spatial visualization ability was added, this step did not explain a significant amount of variance in accuracy, $\chi^2(1)= .04$, R^2 *change*$= .00$, $p= .831$; and spatial visualization ability was not a significant individual predictor, Exp(β)= 1.05, Wald= .04, $p= .831$. This result indicates that spatial visualization ability explains no further variance in accuracy once arithmetic ability is accounted for in computational tasks.

A third step was added to this model to test if the other user variables measured (object visualization ability, expertise as measured by hours, expertise as measured by survey items) and display factors (graphs vs. grables) have any impact on accuracy in computational tasks. These four variables were entered in step three which was significant $\chi^2(4)= 70.08$, $\Delta R^2= .29$, $p< .001$, and resulted in a significant overall model $\chi^2(6)= 76.53$, $R^2= .32$, $p< .001$. The table below shows the results for individual predictors of the model. It was found that only mathematical ability (higher the better), and display type (better accuracy with grables then graphs) were the only significant predictor variables.

Table 14. Logistic Regression; Computational Task Accuracy.

Predictor	β	S.E.	Wald	*df*	*p*	Exp(β)
Mathematical Ability	.24	.10	5.25	1	.022	1.27

Spatial Visualization Ability	.08	.29	.07	1	.791	1.08
Object Visualization Ability	-.13	.29	.21	1	.647	.88
Expertise-survey	.08	.16	.27	1	.604	1.01
Expertise-hours	.01	.04	.04	1	.837	1.00
Display	2.27	.30	57.04	1	<.000	9.67
Constant	-4.84	1.64	8.74	1	.003	.01

Together these regressions show that spatial visualization ability is a significant predictor that explains variance beyond what is captured by mathematical ability for response time but not for accuracy, indicating that in terms of response times spatial visualization ability is an important variable to account for, however ultimate accuracy on tasks involving computations come down to a person's mathematical abilities. Additionally, the third step of the logistic regression helps to test claims of hypotheses 2, 3 and 4—the results of which are discussed in their respective sections.