UNPACKING COGNITIVE BENEFITS OF DISTRIBUTED COMPLEX VISUAL DISPLAYS FOR EXPERT AND NOVICE SCIENTISTS

by

Jooyoung Jang

BA, Yonsei University, Seoul, 2005

MA, Yonsei University, Seoul, 2007

MS, University of Pittsburgh, 2009

Submitted to the Graduate Faculty of

the Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2013

UNIVERSITY OF PITTSBURGH

DEPARTMENT OF PSYCHOLOGY

This dissertation was presented

by

Jooyoung Jang

It was defended on

March 11, 2013

and approved by

Stephen C. Hirtle, School of Information Sciences, University of Pittsburgh

Timothy J. Nokes-Malach, Department of Psychology, University of Pittsburgh

J. Gregory Trafton, Intelligence Systems, Office of Naval Research

Dissertation Advisor: Christian D. Schunn, Department of Psychology, University of Pittsburgh

UNPACKING COGNITIVE BENEFITS OF DISTRIBUTED COMPLEX VISUAL DISPLAYS FOR EXPERT AND NOVICE SCIENTISTS

Jooyoung Jang, PhD

University of Pittsburgh, 2013

The current research focuses on the advantages and disadvantages of two common types of spatially different organizations of information (i.e., spatially stacked vs. distributed) and their impact on science problem solving. The research is based on the premise that we must better understand the spatial organization of information from the perspective of cognitive performance and expertise theories to further our theoretical understanding and provide a practical guide for using and developing effective information visualizations. A new theoretical decomposition and matched analytic technique using eye-tracking is introduced, and is used to tease apart interactions with expertise. Seventy novice scientists and 38 experts participated in the study. They solved a data interpretation problem using either a distributed or a stacked display. Overall, novices took longer to solve the problem when they work with a distributed display than with a stacked display, and eye-tracking data suggests the effect is due to information overload and data management time. By contrast, experts showed a reverse trend (i.e., faster problem solving with distributed displays), being better able to manage complex information. As for the underlying mechanism, three factors (i.e., information internalization, information access, and information externalization costs) were examined and found critical to explain the effect. Both groups showed trade offs among the three factors as an adaptive behavior for effectively balancing the information access costs.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

Psychologists and human factors researchers have argued for the importance of visual display design alike. A vast number of studies found that different visual displays of the same information (i.e., informationally equivalent displays) can yield drastically different task performance because different visual displays are not computationally equivalent (Breslow, Trafton, & Ratwani, 2009; Carpenter & Shah, 1998; Gattis & Holyoak, 1996; Hegarty, Canham, & Fabrikant, 2010; Kroft & Wickens, 2002; Larkin & Simon, 1987; Novick & Catley, 2007; Ratwani, Trafton, & Boehm-Davis, 2008; Sanfey & Hastie, 1998; Shah & Carpenter, 1995; Yeh & Wickens, 2001; Zhang & Norman, 1994). For example, two sets of informationally equivalent graphs were found to be computationally different, and the computational advantages of a new representation could even outweigh the lack of familiarity issues (Peebles & Cheng, 2003). If visual display design has a significant impact for simple displays of small-scale data (e.g., a graph), then the impact should be even greater for complex displays of larger scale data. With technology-based change in visual presentation and complexity of tasks, research on the role of visual display design in complex tasks is essential, as recently noted by Hegarty (2011, p. 450): “developments in information technologies have led to new challenges of how to visualize large and complex data sets with researchers in scientific visualization focusing on displays of spatially distributed data (e.g., the development of a thunderstorm) and researchers in information visualization focusing more on visualization of abstract information spaces (e.g., semantic relations between documents)."

The current research focuses on the advantages and disadvantages of two types of spatial organization and their impact on novice and professional scientists. The research is based on the premise that we must better understand the spatial organization of information from the perspective of cognitive performance and expertise theories to provide a practical guide for using and developing effective information visualizations. Specifically, the research examines (1) the processing of complex visual information of (2) real scientists doing real-world tasks, and thus contributing (3) to extend the cognitive psychology of science and expertise. This study also generates specific implications for the design of hardware and software that scientists use. In short, this research utilizes and extends fundamental cognitive psychology of information processing to both understand and benefit the work of scientists and scientists-in-training.

Many researchers have examined the ways in which users interact with large displays and what performance benefits they gain in using large rather than small displays for the completion of various tasks, such as spatial orientation, reading comprehension, and programming (Grudin, 2001; Mynatt, Igarashi, Edwards, & LaMarca, 1999; Robertson et al., 2005; Tan, Gergle, Scupelli, & Pausch, 2003, 2004). These studies often find that large displays are better (for a review: Czerwinski et al., 2006). However, while many scientists have adopted dual or large screens, other scientists now spend less time using large monitors or have removed desktops entirely from their work and home, relying instead on the smaller screens found in high-powered ultra-light laptops, multi-touch display tablet computers, and Internet-enabled smartphones. Is this move to smaller screens detrimental to complex work like science?

Some research suggests that larger displays are not always beneficial to problem solving. In Kroft and Wickens (2002), pilots-in-training were significantly slower in answering questions exclusively relevant to just one map when using a large integrative display, but faster and more accurate in answering integrative questions (i.e., questions that must be answered by *combining* information from two maps). Critically, the pattern of results suggests that the issue is likely not the size *per se* but rather the spatial organization (see Figure 1)—the observed effects of raw size of display are likely caused by different organizations of information that different display sizes tend to require (Jang & Schunn, 2012; Jang, Schunn, & Nokes, 2011).

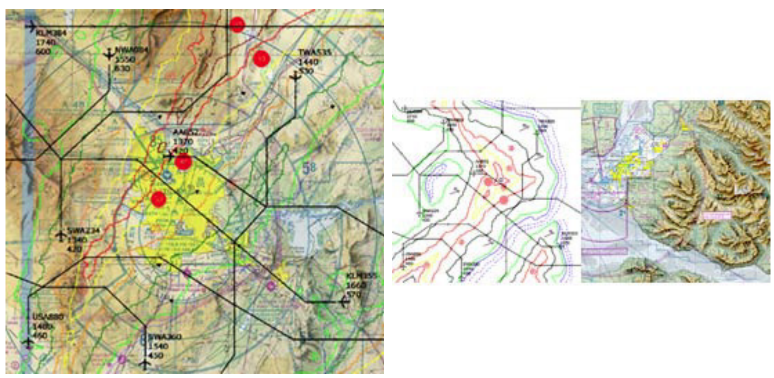


Figure 1. Examples of display maps spatially integrated and separated. From Kroft, P. D., & Wickens, C. D. (2002). Displaying multi-domain graphical database information. *Information Design Journal, 11*(1), p48, Figure 2. Reprinted with kind permission from John Benjamins Publishing Company, Amsterdam/Piladelphia. www.benjamins.com.

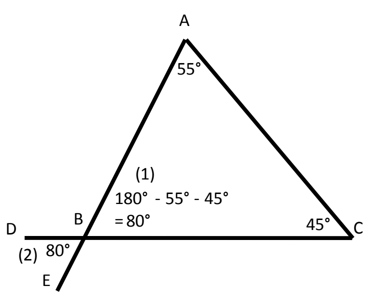
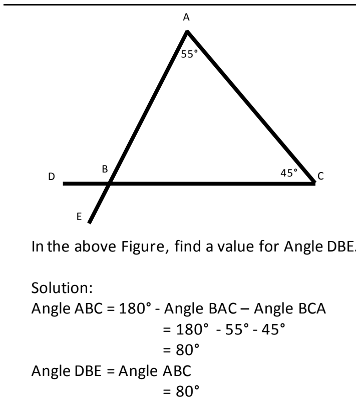
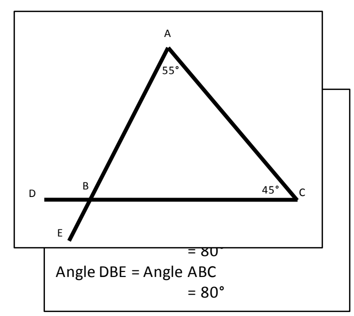
The importance of intelligent use of space, whether in 2D or 3D, was demonstrated in a variety of previous studies (Hollan, Hutchins, & Kirsh, 2000; Kirsh, 1995; Kirsh & Maglio, 1994). Human as "spatially located creatures" are bound to space and effective use of space often reduces time and effort needed to complete a task (Hollan et al., 2000). Also, people often use space as a resource and manipulate spatial arrangements to simplify choice, perception, and internal computation (Kirsh, 1995). Thus, it may be that what is important is not only what data is shown, but also how it is shown to assist timely perception, information integration, and comprehension.

1.1 SPATIAL ORGANIZATION OF VISUALLY PRESENTED INFORMATION

Traditionally, when talking about spatial organization of information displays, the focus of past research was on whether information is presented in an integrative or a separated manner. Integrative displays converge multiple sources of information into a single source (see **Error! Reference source not found.**). For example, when students study a worked example to learn geometry, text instruction can be placed inside the accompanying picture, thus removing the cognitive load required for mental integration. In contrast, separated displays (the middle column) provide to-be-integrated information in disparate spaces and effortful mental integration must precede any learning. Researchers in both education and human factors have argued for the benefits of integrative displays over separated displays for tasks that require information integration (Kroft & Wickens, 2002; Sweller, van Merriënboer, & Paas, 1998).

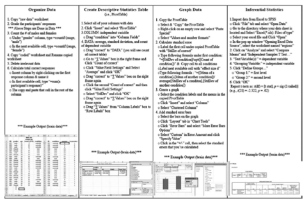
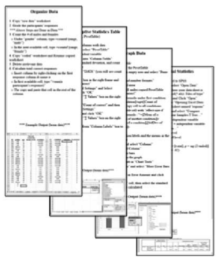
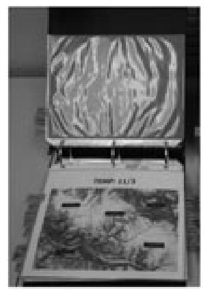
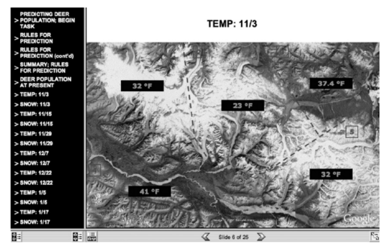
Related to but importantly different from the distinction of integrated vs. distributed displays is another display organization contrast relevant to integrative tasks: spatially distributed displays (i.e., when information sources are presented side-by-side) versus stacked displays (i.e., when information sources are sitting on top of one another with only the top source fully visible). For example, when a meteorologist attempts to make a forecast by integrating information from a large number of maps (e.g., air pressure, wind speed, and cloud distribution maps by the unit time- and height-interval), a single integrative display is not a practical option because superimposing even three such information rich maps would be enough to make the display so cluttered that it would be hard to search and perceive critical information. Because there are so many different information views to examine, the viable options are then to stack the different displays within a small physical space or to distribute the displays across a large physical space.

Table 1. Three types of display organization. The image of a stacked display here presents partially overlapped contents to reveal what was stacked but the two contents should be imagined as fully overlapped.

Integrated	Separated	
	Distributed	Stacked
		

A number of prior studies on this contrast have consistently found large performance benefits of spatially distributed displays over stacked displays across studies in instruction designs and problem-solving domains (Jang & Schunn, 2012; Jang et al., 2011; Jang, Trickett, Schunn, & Trafton, 2012). For example (see Table 2), college students solved integrative problems almost two times faster without any loss of accuracy when information or learning instructions were provided in a distributed format (e.g., 20 information pages printed and pinned on a wall or 4 pages of instructions printed on 11”x17” paper); we have coined this phenomena the distributed display time advantage.

Table 2. Examples of display formats used in prior studies.

Display Type	Jang, Schunn, & Nokes (2011)	Jang & Schunn (2012)	Jang, Trickett, Schunn, & Trafton (2012)
Spatially Distributed			
Spatially Stacked			

1.2 REAL SCIENTISTS DOING REAL-WORLD TASKS

When considering how easy a transition from a stacked to a distributed display is, the time advantage is practically very important. For example, it is a simple matter to switch from one monitor to two, to adopt a large screen that affords viewing multiple windows, or simply to print documents and spread them out on a table. The magnitude of the effect may be even larger in a professional science setting. One study of weather forecasters-in-training (Jang et al., 2012) found a large difference concerning the time forecasters spent to complete a task on the computer (i.e., stacked displays) versus using a map wall (i.e., maps of meteorological information printed out and stuck on a 100” x 40” wall in a distributed display fashion). Even though students with the map wall display could not use animations and map comparison/modeling tools commonly used to improve forecasting, they made predictions 40% faster (25 minutes) than those who used the single monitor computer display (40 minutes), with equal prediction accuracy. Given that rapid and accurate weather prediction is key for work and safety situations that depend upon weather forecasts, the difference is substantial. Interestingly, the study showed that the benefit of a sophisticated computer program based on years of research can be overcome by the consequence of moving to a smaller window size on how the information is spatially organized (distributed or stacked).

The issue of distribution applies well beyond meteorology. In many sciences, a number of highly used statistics packages provide information primarily in a stacked display, which may broadly hinder work efficiency. For example, the Statistical Package for the Social Sciences (SPSS) is the most widely used program for statistical analysis in social science and beyond (e.g., market researchers, health researchers, survey companies, government, education researchers, marketing organizations and others). The original SPSS manual (Nie, Bent, & Hull, 1970) has been described as one of "sociology's most influential books" (Wellman, 1998). Yet, the program provides data tables and graphs in a serial and stacked manner, and it is not hard to meet a researcher who complains about the layout of the output window of SPSS. In the current research, the focus is specifically on real world science problem solving of this type (i.e., data interpretation using statistical information) as a subject matter.

In addition to the practical value, this endeavor is theoretically meaningful as well due to the sparseness of studies that examined the role of visual displays on complex tasks. Most previous studies have focused on simple and well-defined tasks such as extracting specific values, comparing values, or detecting expected trends as the whole task. Although the simple tasks have been very useful for cognitive scientists to characterize the cognitive processes of task performance with various visual displays and develop cognitive models (Carpenter & Shah, 1998; Freedman & Shah, 2002; Shah & Freedman, 2011), as Thomas and Cook (2005) pointed out, these simple tasks are not the most interesting tasks in visual analytics. Real visual analytics work on ill-defined tasks that entail uncertainty and may require data exploration, or reasoning with thousands of data points scattered across multiple visualizations (Ratwani et al., 2008; Trafton et al., 2000; Trafton, Marshall, Mintz, & Trickett, 2002; Trickett, Trafton, Saner, & Schunn, 2007).

Much of science is inherently a complex information integration task. Science work commonly consists of experimenting (whether via thought-experiment, simulation, or physical-experiment), data gathering, analyzing, and interpreting. Due to the inter-dependency of these steps, integration across steps and data sources becomes a crucial and inherent part of science work. The design of an experiment changes the contents of the data that is gathered, and the way the data is analyzed influences data interpretation. Accurate interpretation requires information integration and comprehension across all the steps to draw a sound conclusion and plan a next meaningful experiment. When multiple experiments are examined (as is typically the case), detailed integration must take place across findings, but also to some extent across methods and measures. The step of data interpretation (within and across experiments), a crucial and highly integrative step in conducting science, was used as the focal task for the current research to examine the effect of displays in the context of science problem solving.

Despite the integrative nature of science problem solving and the availability of large display technologies, many scientists often do not make full use of available screen space, and often still rely purely on small laptops that rarely are large enough to accommodate multiple sources of data. For example, we observed this pattern of display use in an observational study (Jang & Schunn, 2008): Scientists from various disciplines were gathered to investigate the nature of water ice below the solid Martian northern arctic plain during the NASA Phoenix Mars '08 Mission. In this setting, all scientists were provided with a high-end computer with a large dual monitor set-up. However, these scientists used only one screen most of the time (63%), and 81% of single-screen use involved a laptop. Here, factors other than display effectiveness likely drove screen use, and this observation argues for the importance of studying experts: are experts less subject to effects of information organization?

To take up this issue of generality of effects across expertise levels, this research examined the effect of spatial organization of scientific information on professional scientists (i.e., experts rather than undergraduates). Many lab studies use simple tasks— easy enough that any untrained person could do—because it simplifies data collection (i.e., all possible individuals can participate and with little-to-no training time). Further, it is often assumed that samples of undergraduates are representative of humans more broadly. However, rather than typical, Jones (2010) characterized undergraduate samples that are usually studied as "WEIRDos". That is, they are people from Western, Educated, Industrialized, Rich, and Democratic cultures," and meta-analyses suggest that the findings from many psychology studies using only typical US undergraduates do not generalize to other populations. As a large segment of the US scientific workforce is not WEIRDos, this issue is particularly problematic for generalizing many existing lab studies on undergraduates to real-world science.

A second generalization issue is the nature of the task. Many 'science' tasks used in psychology labs and science education research are far from what scientists actually do (Chinn & Malhotra, 2002), simply because the pool of undergraduates and young students available for experimentation cannot do more typical science tasks. Further, undergraduates performing science tasks can be very different from scientists doing actual science tasks, just as beginning driver behaviors can be very different from racecar driver behaviors. As a result of these issues, the proposed research directly examines the effects of expertise.

1.3 COGNITIVE PSYCHOLOGY OF INFORMATION PROCESSING AND EXPERTISE

I consider three explanations for the distributed display time advantage: Do stacked displays make people into slow memorizers (information internalization cost), frequent page flippers (information access cost), or note-takers (information externalization cost)? Note that the effect of stacked displays is likely due to a combination of these three explanations.

First, a stacked display may lead to getting lost, or at least more revisiting of information on the path to finding critical information. As demonstrated in Kroft and Wickens (2002), student pilots with the spatially stacked display produced significantly more toggles between the two information sources, compared to those that had integrative displays. Similarly, weather forecasters (Trafton, Trickett, Schunn, & Kirschenbaum, 2007) who worked with a 17-inch desktop revisited maps six times more often than those who had the map wall display (i.e., maps of meteorological information printed out and stuck on a 100” x 40” wall).

Second, in stacked displays, activities such as using notes to keep track of information within and across pages and going back to certain pages may act as additional secondary tasks, even when the task itself is not a dual-task problem. Note taking might be the most commonly used strategy that problem solvers use to avoid information overload and to promote accuracy.

Using notepads has been found to help people solve scientific reasoning problems more accurately and facilitate self-explanation when learning (Trafton & Trickett, 2001). Likewise, stacked display users may frequently take and refer to notes to keep important information externalized.

Third, people may choose to slow down and memorize information to keep it available in their heads rather than keep checking back. Gray and Fu (2004) have shown that people easily become memorizers when the necessary information is even just a click away. Unlike the other two explanations, the slow memorizer explanation provides insights at the level of the underlying cognitive mechanism (i.e., strategy selection depending on the degree of information access cost that a display imposes) and systematically influences the other two explanations (i.e., memorizers should turn pages and get lost less often than verifiers). The next paragraph presents data that supports the slow memorizer explanation and how this single explanation can be consistently applied to different studies.

1.3.1 Information Access Cost and Memorization

A recent eye-tracking study suggested that problem-solvers adopt an information memorization strategy in stacked conditions, and this memorization time could account for the stacked display time disadvantage (Jang et al., 2012). Eye tracking involves measuring the point of a gaze (called a fixation) and the length of time the gaze remains fixed on a location (called a fixation duration). Participants in the stacked display condition fixated significantly longer on information pieces on each page throughout an integrative problem-solving task than those who solved the same problem using the distributed display (see Figure 2), as a (possibly unconscious) strategy to bypass the relatively higher information access cost in the stacked display. That is, the stacked display produces a situation with high information access costs because information is a page-turn away, compared to the cost of an eye/head turn in the distributed display. Consequently, problem-solvers chose to memorize information (by staring at each piece of information a little longer) rather than repeatedly turn pages to look for information.

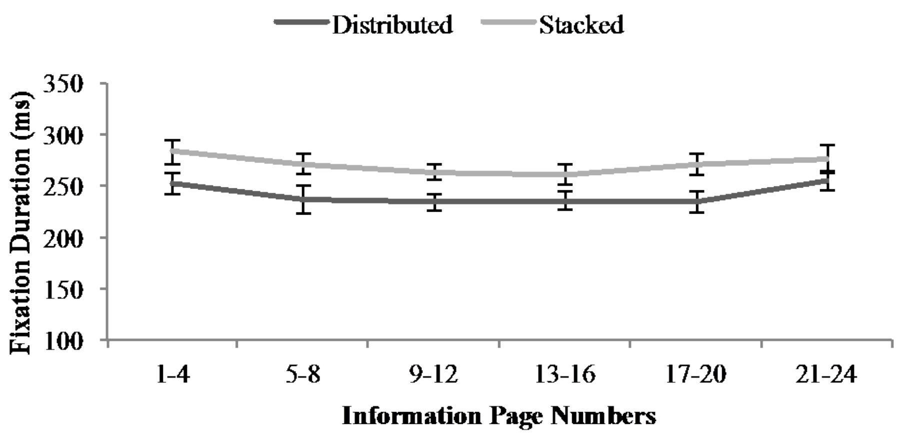


Figure 2. Average first pass fixation duration on information presented in distributed vs. stacked formats

When the cost of accessing external information increases, people tend to memorize information to make it readily accessible *in the head* (i.e., memorization strategy; stacked display). By contrast, when information access cost is low, people do not bother to memorize information and instead rely on external/*in the world* information (i.e., perceptual-motor strategy; distributed display). In terms of performance accuracy, the memory strategy selection can be construed as an adaptive choice balancing accuracy and effort, because information *in the world* is accurate but that *in the head* may not be. For example, participants made more errors in a given task when they adopted the memorization strategy, but with a reduction in task time (Gray & Fu, 2004). While in science we would prefer that scientists not trade accuracy for speed, there will also be some tradeoff because the set of possible analyses is infinite and time is finite.

One may argue that the slow memorizer explanation goes against the student weather forecasters study (Jang et al., 2012) where people in the stacked display condition revisited pages far more frequently, which should not have occurred if they had memorized information. In fact, the eye-tracking study and the student weather forecaster study together demonstrate the power of the adaptive choice theory underlying the slow memorizer explanation. The slow memorizer explanation suggests that differing information access costs associated with information layout affect the probabilities of adopting different information encoding strategies. The stacked view display makes it harder to access information that needs to be integrated; one can compensate either by slowing down and memorizing information as observed in the eye-tracking study or by frequently going back and forth as in the weather study. The choice of strategy likely can be explained by relative access costs: hovering vs. clicking. People in the weather interface likely became verifiers because the interface made flipping maps very easy through hovering (simply holding the mouse over different areas changes the animation content immediately). By contrast, people in the eye-tracking study became memorizers, as they had to click and wait hundreds of milliseconds to access a content page. However, regardless of which strategy was used, the stacked display costs users a substantial amount of total problem solving time that could have been saved by using a distributed display.

1.3.2 Expertise Effects and Expertise Reversals

It is unknown whether the distributed display advantage phenomenon and the slow memorizer explanation of it would hold up in the case of experts. They may hold up in the expert case because the information access cost is a basic factor that affects all human information perception (i.e., low-level cognition). With extensive training and a vast amount of knowledge, however, experts may overcome the costs in ways that non-experts cannot, such as with unique memorization strategies, data monitoring techniques and information search strategies (i.e., high-level cognition).

Likewise, information display research has been focused more on novices such as student pilots, electrician apprentices and trainees, or middle school students. What aids novices may not generalize to experts or may even slow experts. Several studies investigating the relationship between the level of domain knowledge and instructional design (Kalyuga, Chandler, & Sweller, 1998; McNamara, Kintsch, Songer, & Kintsch, 1996; Yeung, Jin, & Sweller, 1998) suggested a so-called expertise reversal effect. That is, advanced learners (i.e., intermediate electricians) learned less when information was provided in integrated texts and diagrams and learned more from a diagram-only format (Kalyuga et al., 1998). Separated displays produce an additional step of mental integration of information pieces for novices and are thus harmful, but for advanced learners, integrated displays provide redundant, distracting information that cannot be ignored and thus hinders learning. A similar reversal of what is best for experts vs. novices was reported in medical diagnoses. While residents performed well with verbal descriptions of dermatology lesions and photographs, professional dermatologists performed worst with verbal descriptions and best with a photograph alone (Kulatanga-Moruzi, Brooks, & Norman, 2004).

Although similar predictions are seemingly possible for the case of distributed vs. stacked displays, it is unclear whether the expertise reversal effect would be also observed in our case. Unlike integrated displays, distributed displays always keep the multiple sources of information as separate entities. Thus, even though the sources may be placed close to each other, each source can be considered as a chunk and more easily disregarded when necessary. By contrast, the inhibition of redundant and distracting information is only one of the several routes that could cause an expertise reversal effect. For example, expert chess players can carry out several games simultaneously without the view of the board and the pieces in a game called blindfolded chess. They can monitor, update, remember, and reconstruct chessboards of more than 50 moves in their head with only a 10% or less error rate (Saariluoma & Kalakoski, 1997), and they are even able to generate best moves in given problem situations (Saariluoma & Kalakoski, 1998). These amazing abilities depend upon their *eyes* for search and predicting relevant moves and from abundant yet flexible scenarios established from their vast array of prior experiences.

1.3.3 Summary

To summarize, the three goals of the research were to: (1) examine the effect of information display (i.e., spatially distributed vs. stacked) in the context of a real science problem using data interpretation as a complex problem-solving task, (2) examine the effect of information display across populations differing in expertise, which would directly compare novices (i.e., undergrads) and experts (i.e., graduates and post-doctorates, and (3) build an empirically-based framework of the underlying cognitive mechanisms.

To clarify how the multiple factors could operate together, I describe the underlying mechanisms in a diagram (see Figure 3) that is proposed in this study. In this framework, I examined the cognitive process of information integration in complex problem solving considering fundamental aspects of information processing that likely do not change with expertise and those more experience dependent aspects: how differing information access costs of the two information display formats affect the information encoding process (e.g., how experts' performance is constrained by information displays) and how expertise affects information management strategies such as note-taking strategies in response to a given information display (e.g., how experts use their skill sets and knowledge to effectively overcome the constraints an information display imposes).

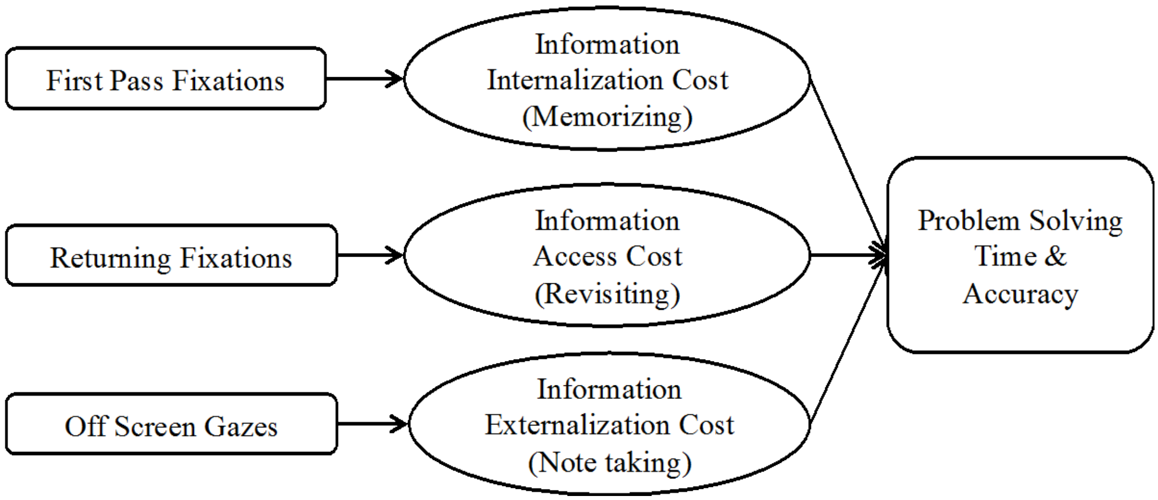


Figure 3. A framework of underlying mechanisms

There are three types of cognitive costs in relation to information access efforts: information internalization cost (i.e., memorizing measured by first time fixations), information access cost (i.e., revisiting measured by returning fixations), and information externalization costs (i.e., note taking measured by off screen gazes). Systematic research on the composite effect of these factors has not been previously done. I propose that these factors would together explain task performance well, and expertise would moderate the underlying cognitive processes.

2.0 EXPERIMENT: NOVICES VS. EXPERTS

2.1 METHOD

Scientists from two different points on the expertise continuum (i.e., psychology undergraduates vs. psychology graduate students and post-docs) were given a data interpretation task in two display formats (i.e., distributed vs. stacked) and their performance was compared to examine possible moderation of expertise on the effects of display format on performance. Eye-tracking data was used to examine possible underlying mechanisms.

2.1.1 Participants

The novice group consisted of 70 psychology undergraduate students (35 female; age range 18-25) at the University of Pittsburgh participating for course credit; the students were primarily 3rd and 4th year students who had taken a number of psychology content courses already and had some coursework in psychology research methods. The expert group consisted of 38 psychology graduate students and two post-docs (32 female; age range 23-48) at the University of Pittsburgh and Carnegie Mellon University participating for $25.

To ensure that expert participants had relevant expertise in data analysis and interpretation, several expertise measures were examined. On average, expert participants were in their 4th year of a PhD program (*SD* = 2.2) and had 4.7 years of experience with behavioral data (*SD* = 3.0) and 4.5 years of experience with ANOVA (*SD* = 2.9). Further, expert participants had a mean of 2.7 journal publications (*SD* = 3.9) and 2.8 additional non-journal publications (e.g., conference papers, *SD* = 3.0). Of these, a mean of 1.1 were first author publications (*SD* = 2.5), and 1.2 were papers published in the prior two years (*SD* = 1.2). That is, the participants typically had multiple years of successful research experience and were currently research active.

Expert participants were randomly assigned to either display format, and there were no differences in the level of expertise between experts in each display condition (see Table 3).

Table 3. Mean and SD of expertise measures within each display format condition

	Stacked		Distributed		
Measure	M	SD	M	SD	*p* value
Year in PhD program	4.2	2.4	3.9	2.2	.72
Years of experience with Behavioral data	4.6	3.5	4.8	2.5	.79
Years of experience with ANOVA	4.3	3.3	4.6	2.6	.75
Number of journal publications	2.6	2.5	2.8	5.0	.87
Number of non-journal publications	2.3	2.5	3.4	3.4	.26
Number of first author publications	0.7	0.9	1.6	3.4	.28
Number of publications in last 2 years	1.2	1.3	1.1	1.1	.79

2.1.2 Design

A 2x2 between-subject design was used. The independent variables were display format (distributed vs. stacked) and expertise level (novices vs. experts). In individual sessions, both the novices and experts solved a data interpretation problem using either a distributed or a stacked display. The task was self-paced to measure efficiency as well as accuracy of problem solving using distributed vs. stacked displays. While novices and experts solve the task, problem solving time, page transition logs, and eye movements were recorded. A set of survey was collected at the end of the experiment. The dependent variables include problem-solving times (i.e., time on information window, time on question-answering, and total task time), task accuracy, number of page visits, and patterns of eye-movements.

2.1.3 Materials

2.1.3.1 Eye-tracker.

Eye-movements were recorded with a Tobii 1750 remote eye-tracker. The 17" monitor's screen resolution was 1280 x 1024. The system runs at a constant frame-rate of 50 Hz. The approximate distance between the screen and participant was 12".

2.1.3.2 Main task materials.

The main task involved quantitative data interpretation, examining research hypotheses with quantitative data presented in tables and graphs, and drawing a plausible conclusion. The characteristic of quantitative data interpretation of particular relevance to manipulations of display format is its integrative nature. Data interpretation is an integrative task because it involves processing multiple pieces of information (e.g., dependent and independent variables; descriptive and inferential statistics) presented in various formats (e.g., text, tables, graphs, and diagrams) and combining the information into a single coherent story. It is also a complex task that requires critical thinking. For accurate data interpretation, one needs to understand how the research was conducted (e.g., knowing strengths and weaknesses of the design), examine statistical information critically (e.g., thinking about which data support or contradict the tested hypotheses, which data are more convincing than others, and what additional analyses can be done), and make plausible inferences in relation to research hypotheses (e.g., drawing a conclusion regarding which hypothesis is supported based on available evidence and how other contradicting results can be explained in the frame of the chosen hypothesis).

To avoid the complicating effects of domain specific knowledge across differing focal areas of expertise among the expert group (e.g., cognitive, developmental, clinical psychology, or neuroscience), the task materials were consist of phenomena that are understandable to a broad range of research psychologists (Schunn & Anderson, 1999). The topic and content of the data was selected from *Psychological Science*, a psychology journal that delivers brief research reports of broad interest, so as to make support comprehension and interest in the task in the novices. Real journal reports were used to maintain plausibility of the task for experts. The particular topic involved destination memory (e.g., remembering the person to whom one has given information) and source memory (e.g., remembering the person from whom one has received information), examining which memory is more fallible and why. The content of the task data was adapted from Gopie and MacLeod (2009), Koriat, Ben-Zur, and Druch (1991), and Marsh and Hicks (2002).

The materials involved a brief description of the research topic, data from two studies that each provide evidence consistent with one hypothesis but contradict each other. A one-page paper handout motivated the general research questions and provided participants with definitions of key concepts such as destination and source memory. On the computer, 13 content pages were available: questions to be answered, study 1 intro, study 1 hypothesis, study 1 methods (1) and (2), study 1 results, study 2 intro, study 2 hypothesis, study 2 methods (1) and (2), study 2 results (1), (2), and (3).

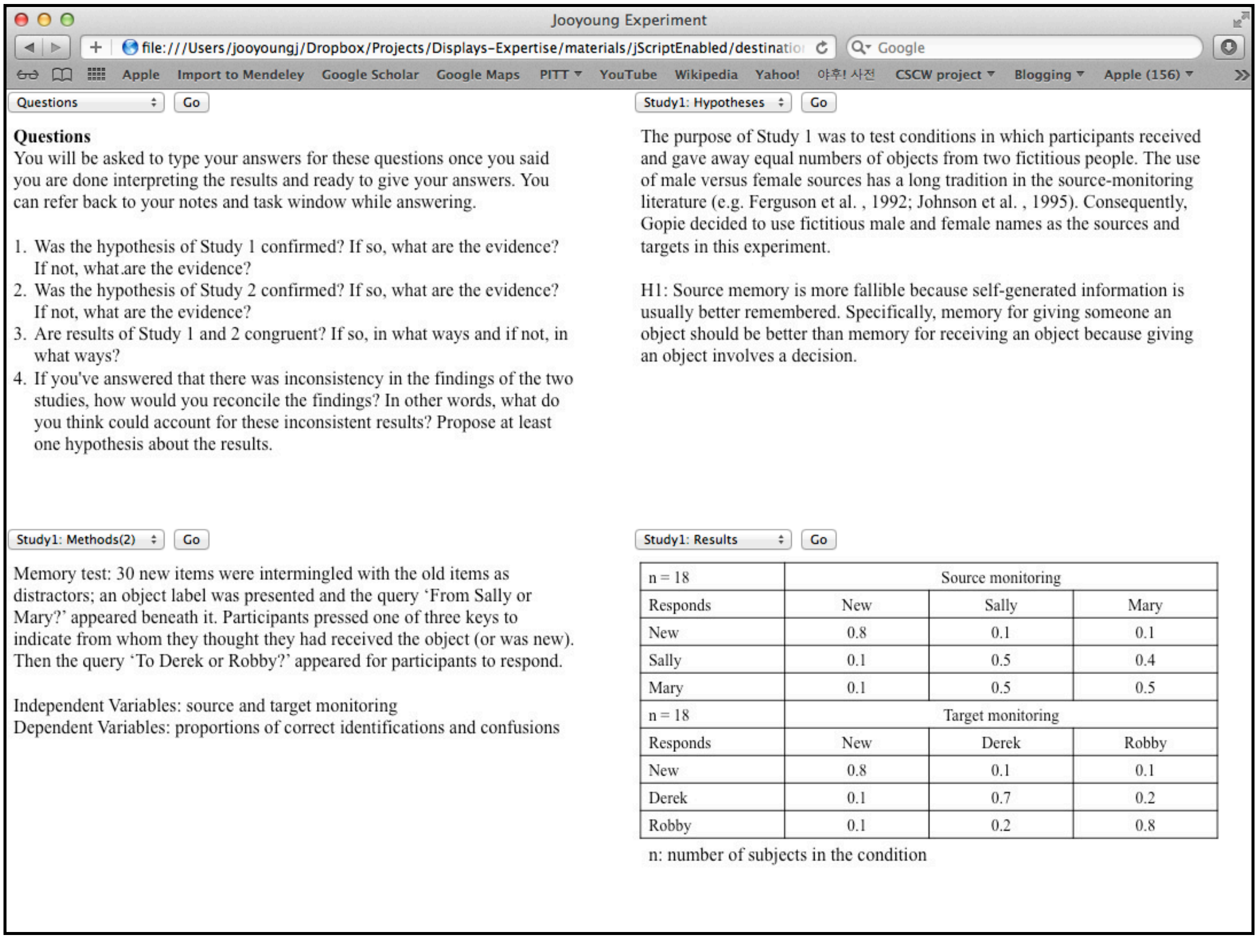


Figure 4. Layout of spatially distributed display in the current experiment

In this study, the two formats of information displays were defined in the following manner. The distributed displays divided the 17-inch screen into four equal-sized spaces (see Figure 4). Each space had a drop-down menu with which participants could choose the information that was loaded into each space. This was the same set-up that was used in the prior eye-tracking study (Jang et al., 2012) except that this time, participants in the distributed display condition can manage each information window separately. Thus each participant determines which information page to view in which space, resulting in unfixed and user-dependent sets of information page combinations.

By contrast, the stacked displays present every information page in the top-left one-quarter of the 17-inch screen space, leaving the other three spaces blank. Thus, only one information page could be examined at a time in the staked displays, and information shown previously was replaced by the next chosen display. The current set-up of stacked and distributed display simulates the use of small vs. large screens while keeping screen resolution *per se* constant. In both display format conditions, the screen was initially populated with blank pages.

2.1.3.3 Practice materials.

A short and simple practice was developed to familiarize participants with the general procedure of the task. The topic of the practice was learning with diagrams (why and when having a diagram improves learning) and seven pages of information were provided. The content of practice data was adapted from Willows (1978).

2.1.3.4 Surveys.

Three surveys were developed: strategy, cognitive load, and demographic surveys. First, the strategy survey consisted of questions regarding perceived information availability and explicit use of a memorization strategy (i.e., whether they tried to memorize information), and other possible strategies they may have used to solve the task. Responses were made on a typical five-level Likert scale ranging from ‘Disagree’ to ‘Agree’.

Second, the cognitive load survey asked participants to rate perceived difficulty and the amount of mental effort invested on 9-point scales, ranging from very, very easy (1) to very, very difficult (9) and from very, very low mental effort (1) to very, very high mental effort (9). This subjective rating scale has well-documented validity and reliability (Paas & van Merriënboer, 1994). It was expected to show the differential load participants perceived with the two display formats. In our previous study, a higher cognitive load was reported for stacked displays (Jang et al., 2011).

Lastly, a demographic survey was employed to gather background information on age and gender. For the expert group, additional items were included: PhD program within psychology, year in the program, years of experience with behavioral data, and years of experience with ANOVA (the type of data analysis presented in the main task).

2.1.4 Procedure

The experiment was done in individual sessions and the session consisted of three components: practice, task, and survey. Each participant was first seated at an eye-tracker (Tobii 1750) with a chin rest. After a brief eye-calibration, participants performed a practice task in the stacked display format. Before the practice started, participants first read aloud a passage describing the topic of the practice problem and confirmed that they understood the topic (see Appendix A). Then they were given instructions on how to navigate information pages using a drop-down menu, what are the questions they need to solve, and that the information window will be available at all times. Also, they were instructed to let the experimenter know when they are done with information window and ready to compose the answer so that they can move to another workstation equipped with a keyboard. A blank letter-size paper was provided to each participant to take notes during the problem solving. Participants were asked to make sure that their chin is on the chin-rest while they examine information on the screen but were allowed to lean back while they take notes. After the instruction, participants were asked to try the practice problem alone (see Appendix B). During the practice, they were allowed to ask questions either about the problem content and the procedure. For the practice task, participants were not actually asked to write their answers in order to keep the practice brief, but they were not given this instruction until they said they are done examining the information window. Once they notified the experimenter that they are ready to give their answers, they were guided to the task.

The task (see Appendix D) was done in the same manner as the practice was conducted. Before the task, participants read aloud a passage describing the topic (see Appendix C) and then were informed that they would write the answers for real this time and the task would be harder than the practice. Those who participated in the distributed display condition were given additional instructions on how to use the distributed display; a sample distributed information window was shown to participants using the practice material. A new blank paper was provided for note taking. Unlike practice, participants were not allowed to ask questions about the task content. When they indicated that they are finished with information examination on the window, eye movement recording was stopped and saved. Then, on another workstation equipped with a keyboard, participants were provided with a new page that has two windows accessible via clicking tabs, one to type their answers under each question (see Appendix E) and the other to hold the information window. Also in a self-paced manner, participants composed their answers for each and every question. They were able to refer back to the information window and their notes as they write.

The goal of the task was to compose short paragraphs for several key integrative questions: whether the hypothesis of the first study was confirmed and why, whether the hypothesis of the second study was confirmed and why, whether the two studies are congruent, and how to reconcile the discrepancies if they are not congruent. The core process of problem solving for the data interpretation task is provided in Figure 5.

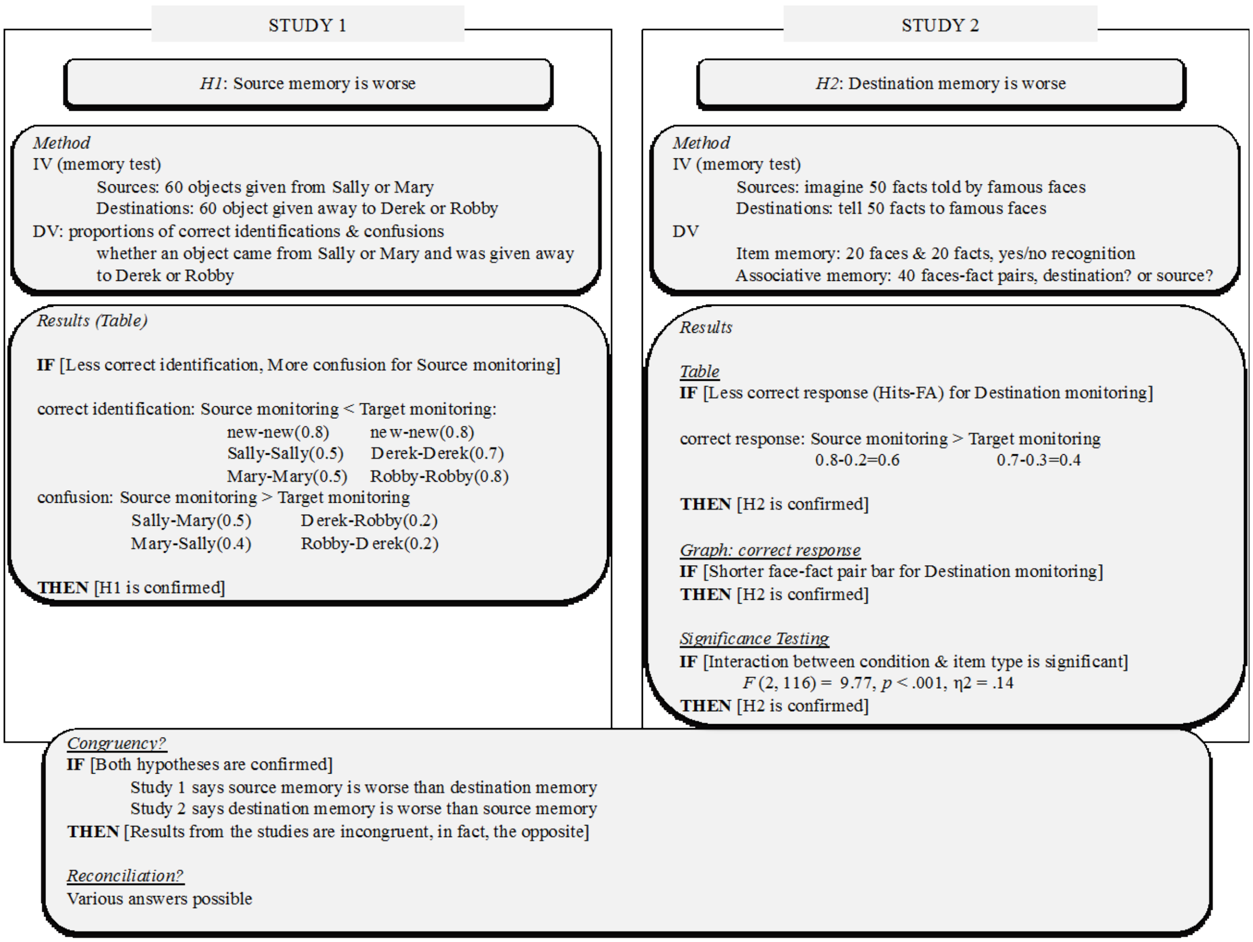


Figure 5. A flow chart of problem solving process for the current data interpretation task. The IF-THEN statements denote critical information that needs to be captured by problem solvers to achieve accurate data interpretation. This chart is provided solely for readers' convenience; it was never shown to participants.

After the task was done, participants completed the three surveys in the order of strategy (see Appendix F), cognitive load (see Appendix G), and demographic survey (see Appendix H). Expert participants were additionally asked to submit their curriculum vitae. Participants were asked to use a given codename instead of a real name when completing task questions and surveys.

2.2 RESULTS AND DISCUSSION

Outliers were defined as instruction violation and total task time longer than two standard deviations from the mean, which could be signs of misunderstanding of the experiment, insufficient basic knowledge to solve the task, and very low task motivation. Six undergraduate participants in the novice group were excluded from analyses as outliers, which left 34 participants in the distributed condition (24 females) and 30 in the stacked condition (19 females). Due to eye-tracker malfunction, three additional undergraduates (two in distributed and one in stacked) were excluded from eye movement data analyses. There were two outliers in the expert group whose total task time was longer than two standard deviations from the mean, leaving 19 experts in each condition (15 females in each condition).

2.2.1 Task Time and Accuracy

Three time measures were available: time spent solving the practice problem (practice time), time spent processing information presented in a given window display (window time), and time spent composing answers while having the information window available on the second tab of a window (answering time).

Expert participants in the two conditions did not differ in practice time, t (36) = 1.19, p = .24 but novice participants in the distributed condition took longer (M = 6.1 minutes, SD = 2.0) to finish the practice task than participants in the stacked condition (*M* = 5.1 minutes, *SD* = 1.0), t (62) = -2.55, $p < .01$, Cohen's $d = 0.68$. To control for general task speed differences across participants, 2x2 MANCOVA was used to test the effect of display format and expertise level on window time, answering time, and task accuracy, with practice time as a covariate. Descriptive statistics are presented in Table 4. More detailed statistics are presented in Appendix I.

Table 4. Mean (and standard deviations) window time, answering time, and task accuracy by display format and expertise level

Expertise level	Display format	
	Stacked	Distributed
	Window time (minutes)	
Novice	13.6 (4.0)	19.2 (8.2)
Expert	20.1 (9.3)	15.7 (4.5)
	Answering time (minutes)	
Novice	12.1 (5.3)	11.5 (4.8)
Expert	17.5 (8.6)	17.1 (6.4)
	Task accuracy (percentages)	
Novice	40.0 (26.7)	27.2 (22.5)
Expert	44.7 (35.9)	56.6 (28.7)

The window time was not different by the display formats, F (1, 97) = 0.01, *MSE* = 0.14, $p = .95$, nor by the expertise level, F (1, 97) = 0.65, *MSE* = 21.23, $p = .42$. More importantly, however, there was a significant interaction effect, F (1, 97) = 6.55, *MSE* = 212.69, $p < .01$, $\eta^2 = 0.06$. As shown in Figure 6 (left), the benefit of display formats was differed by expertise level. Experts who used distributed display tended to finish examining information faster than experts who used stacked display (t (36) = 1.87, $p = .07$). But for novices, the effect was the opposite;

novices finished examining information faster when they used the stacked display than when they used the distributed display (t (62) = -3.43, $p < .001$, Cohen's $d = 0.92$).

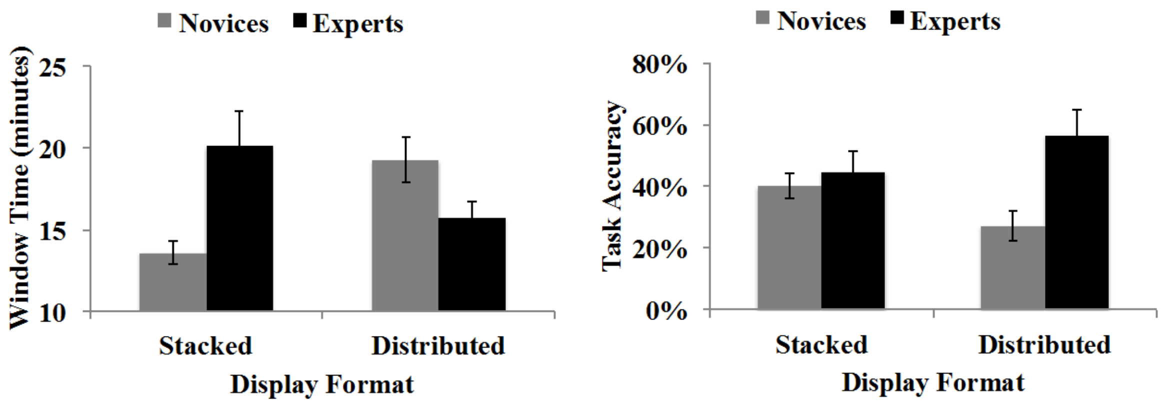


Figure 6. Mean (with SE bars) window time (left) and task accuracy (right) by display format and expertise level.

The answering time was different by expertise level (see Table 4), F (1, 97) = 18.74, MSE = 704.10, $p < .001$, $\eta^2 = 0.16$. On average, experts spent 5.5 minutes more composing their answer than novices did, presumably, from engaging in more thinking and generated more complete and accurate answers; accuracy effects are explored below. This general expert-related slowing held regardless of condition: there was neither an effect of display format, F (1, 97) = 0.17, MSE = 6.47, p = .68, nor an interaction effect, F (1, 97) = 0.02, MSE = 0.80, p = .88 for answering time.

Accuracy was scored according to the grading rubric presented in Table 5. A point was given to each statement that is similar to the answer key and the aggregated points were used as task accuracy. Note that these answers can only be generated from inferential and integrative thinking by making comparisons across pages of hypotheses, methods, and results sections. Statistics for task accuracy are provided in percentages.

Table 5. Grading rubric for the data interpretation task.

Questions	Examples of key answers
Question 1: Was the hypothesis of Study 1 confirmed? If so, what are the evidences? If not, what are the evidences? Be specific.	• Destination memory was better than source memory (more correct identification) • Source memory was worse than target memory
Question 2: Was the hypothesis of Study 2 confirmed? If so, what are the evidences? If not, what are the evidences? Be specific.	• Destination memory was worse than source memory (low correct recognition) • Source memory was better than target memory
Question 3: Are the results of Study 1 and 2 congruent? If so, in what ways and if not, in what ways?	• In Study 1, source memory was worse than destination memory but in Study 2, destination memory was worse than source memory
Question 4: If you've answered that there was an inconsistency in the findings of the two studies, how would you reconcile the findings? In other words, what do you think could account for these inconsistent results? Propose at least one hypothesis about the results.	• In Study 1, the participant makes a decision regarding the destination of the information, while in Study 2 the participant does not; the self-generated information/actions might make destination memory superior under certain circumstances

As expected, experts were more accurate (M = 50.7, SD = 32.6) than novices (M = 33.2, SD = 25.2), F (1, 97) = 8.89, MSE = 6917.48, $p < .01$, $\eta^2 = 0.08$, confirming that experts did have more knowledge and skills than novices for the given task. There was no main effect of display format on task accuracy, F (1, 97) = 0.01, MSE = 4.13, p = .94. But, interestingly, there was an interaction effect, F (1, 97) = 4.23, MSE = 3289.26, $p < .05$, $\eta^2 = 0.04$. Experts in the distributed display condition tended to be more accurate than experts in the stacked display condition, but novices in the distributed display condition tended to be less accurate than novices in the stacked display condition (see Figure 6 right).

One could argue that there can be a speed/accuracy trade off since experts spent more time on composing answers than novices. However, it is unlikely because of the following reasons. Experts in the stacked display condition scored considerably low in the stacked condition, thus even though experts spent more time answering questions (regardless of the display format) than novices, they did not necessarily score higher. In addition, experts in the distributed display condition scored higher than experts in the stacked display condition even though the two expert groups spent a similar amount of time answering questions.

The observation that experts scored only slightly better than novices in the stacked display condition may bear two explanations. First, experts who participated in this study were graduate students in training, thus they may not necessarily possess far superior expertise than undergraduate majors. The finding that experts scored only about 50% on average in total across the display format conditions can also support this idea. Second, the stacked format might have limited the use of expertise by providing an overly simplified problem space. When considering how important it is to examine the big picture in scientific problem solving, compared to the distributed display format, the stacked display format may constrain even (near) experts in developing broad perspective.

In sum, the results show very different effects of display format for experts than for novices: a complete expertise reversal for both time and accuracy. Specifically, a distributed display is beneficial for experts but a stacked display is beneficial for novices.

In the following sections, three major contributing factors (i.e., internalization of information, external information access cost, and externalization of information) that might explain the differences in window time are analyzed.

2.2.2 First Pass Fixation Durations: Internalization of information

Generally, two eye movement measures (i.e., number of fixations and mean of fixation durations) are useful to test the memorization strategy hypothesis. The number of fixations directly shows how many times people fixated on information and the mean of fixations demonstrates for how long people fixated on each information piece, thus the large number of fixations and the longer duration of mean fixations would mean more effortful encoding or memorization. Particularly, the mean of fixation durations was used in a previous study that successfully showed the memorization strategy effect in a stacked display format (Jang et al., 2012). In the current study, however, the sum of fixation durations (i.e., a composite measure that reflects both the number of fixations and the average fixation durations; sum = number x average) was used because it was a measure at the appropriate level in order to effectively test the contributions across first pass fixations, return fixations, and off gaze durations.

To examine the possible role of a memorization strategy in producing the time differences between distributed vs. stacked displays, the sum of first pass fixation durations were computed by summing up the durations of fixations that laid upon each page during the first visit, excluding any regressions or returns. The sum of first pass fixation durations can be a direct measure of the memorization hypothesis, because each fixation duration serves as an on-line measure of information processing, similar to the eye-mind assumption and immediacy assumption used in eye-tracking studies of reading processes (Carpenter & Just, 1983). By examining the sum of first pass fixation durations, one can estimate to what extent a memorization strategy effect contributes to the total time differences observed on window time.

Importantly, based on previous results using eye-tracking to study display format effects (Jang et al., 2012), longer first pass fixation durations in the stacked display condition were expected, which would be consistent with the adoption of a memorization strategy due to the relatively high information access cost of the stacked view display (i.e., having to search and click an index, which involves at least several hundred of milliseconds extra for each information page visit). That is, if problem solvers experience higher information access cost when using stacked displays, they should try to overcome the cost by spending extra encoding time to facilitate later retrievals from memory (Morgan, Patrick, Waldron, King, & Patrick, 2009).

Further, no interaction was expected between the display format and the expertise level under the assumption that the differential information access cost imposed by the display format is a *physical* constraint. The cost of information access was fixed by the display format (i.e., an eye or head turn away in distributed displays and a page-turn away in stacked displays) and thus cannot be modified through expertise, so stacked display users should choose an effortful information encoding strategy (i.e., memorization) to circumvent the extra access cost regardless of their expertise level.

A 2x2 ANOVA was used to test the effect of the display format and the expertise level on the sum of first pass fixation durations. Detailed statistics are presented in Appendix J. As predicted, there was no main effect of expertise level, F (1, 95) = 1.03, MSE = 4.56, p = .31 and no interaction effect, F (1, 95) = 0.01, MSE = 0.04, p = .93. There was only a main effect of display format, F (1, 95) = 98.95, MSE = 438.97, $p < .001$, $\eta^2 = 0.51$. Regardless of the expertise level, stacked display users more than tripled the time initially encoded information (M = 6.4 minutes, SD = 2.5) than distributed display users did (M = 2.0 minutes, SD = 1.6), which supports the memorization strategy hypothesis.

The memorization response to changing access costs should also be seen throughout the information pages of the task. To further whether the effect was throughout the task or localized to particular pages, the sum of first pass fixation durations was analyzed by information page (see Figure 7). As predicted by the memorization strategy hypothesis, both experts and novices in the stacked display condition showed much longer first pass fixation durations consistently across all 13 information pages.

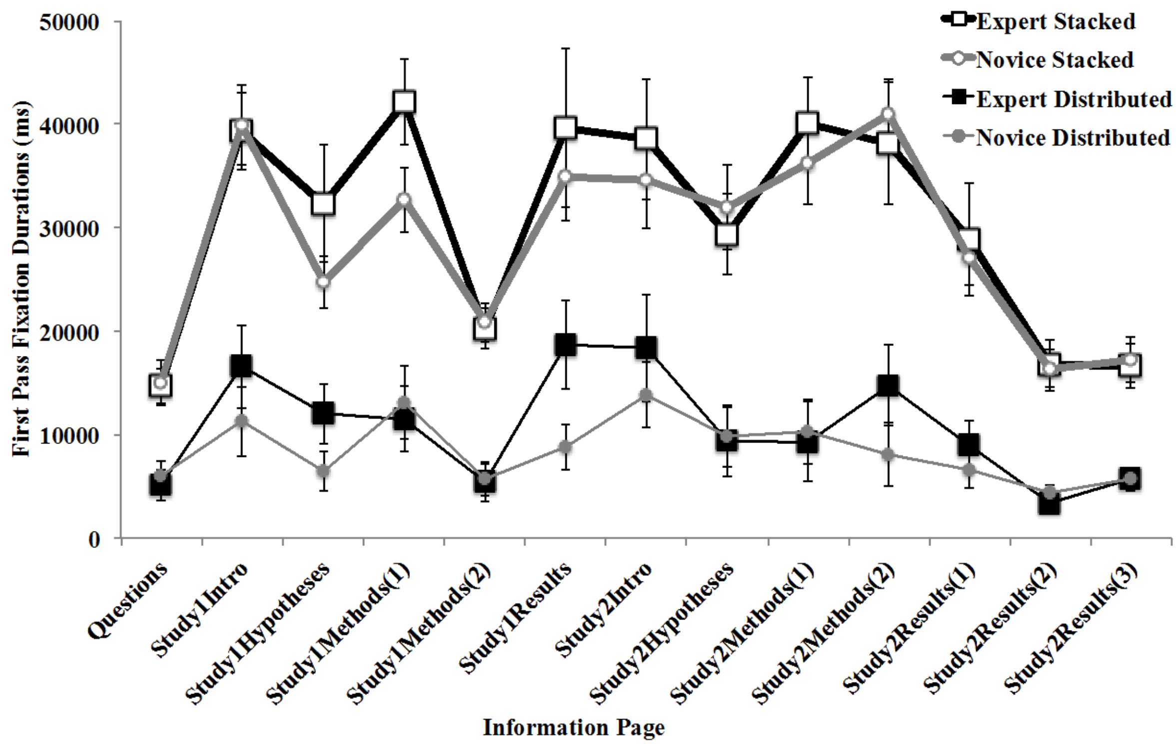


Figure 7. Mean sum of first pass fixation durations by information page for each condition (with SE bars).

To examine the relative extent to which the expertise and display formats might have affected the process, correlations of the profile across conditions were analyzed. By comparing the magnitudes of correlations between the same formats (stacked-stacked and distributed-distributed) to those between the same expertise levels (novices-novices and experts-experts), one can infer whether or not information layout or background knowledge influences the performance profile more for novices and experts. The correlations of the profile across conditions showed a strong stacked-stacked correlation (r = .93, p < .001, n = 13) and a medium distributed-distributed correlation (r = .70, p < .01, n = 13), consistent with the interpretation that the stacked condition was more constraining on expertise effects. Interestingly, for novices, the effect was driven more by the display format than expertise (rather weak novice-novice correlation compared to the strong stacked-stacked correlation: $r = .75$, $p < .001$, n = 13). But for experts, the effect was driven more by their knowledge-based reactions to the input than the display format (relatively strong expert-expert correlation compared to the medium distributed-distributed correlation: $r = .84$, $p < .001$, n = 13).

2.2.3 Return Fixation Durations: External information access cost

The sum of return fixation durations was computed by subtracting the sum of first pass fixation durations from the sum of total fixation durations, which includes all regressions and returns. While measures of first pass fixations show how much time and effort problem solvers invested during the first time encoding in reaction to the relatively higher information access cost in the stacked display condition, return fixations provide an index of external information access cost. Note that total external information access cost involves planning returns, mouse movements, eye-saccades, in addition to the fixation duration of the return. Those other elements are harder to capturing from the eye data. Because those other elements are a multiple of the return fixations, I focus on the return fixations as the tip of the iceberg. In addition, from watching replays of the eye-data, it appears that the bulk of the time is spent on revisiting the information rather than getting back to the information (i.e., return fixations are the bulk of the external information access cost).

If the distributed display produces relatively lower information access cost (i.e., an eye-turn) than the stacked display, distributed display users should not bother to memorize information at first time encoding and therefore make more regressions and revisits. Thus, it was predicted that stacked display users would show shorter return fixation durations than distributed display users. Neither an effect of expertise level nor an interaction between the display format and the expertise level were expected since the effect of return fixation durations relies on the physical constraint as the effect of first pass fixations does.

A 2x2 ANOVA was used to test the effect of display format and expertise level on the sum of return fixation durations. Detailed statistics are presented in Appendix K. As predicted, there was no significant effect of expertise level on the sum of return fixation durations, F (1, 95) = 0.02, MSE = 0.15, p = .90 and no interaction effect, F (1, 95) = 1.41, MSE = 12.38, p = .24. There was only an effect of display format, F (1, 95) = 70.60, MSE = 620.04, p < .001, η^2 = 0.43. Regardless of the expertise level, stacked display users re-accessed information for much less total time (M = 2.1 minutes, SD = 1.6) than distributed display users did (M = 7.4 minutes, SD = 3.8), which is also consistent with a relatively higher information access cost in stacked display condition and further supports the memorization hypothesis as an explanation of the effects of display format.

To further examine the relative localization of this effect across the task, the sum of return fixation durations was analyzed by information page (see Figure 8). Note that for the stacked display condition, these return fixations are purely *revisit* fixations (i.e., a now hidden prior page must be found and re-displayed); these revisits therefore show the pages participants were willing to revisit in spite of the relatively high information access cost. The correlations of the profile across conditions showed a strong stacked-stacked correlation (r = .88, p < .001, n = 13) and a weak distributed-distributed correlation (r = .44, p = .14, n = 13). It appears that for both novices and experts, the effect was driven more by the display format than expertise (rather weak novice-novice correlation compared to the strong stacked-stacked correlation: r = .69, p < .01, n = 13 and expert-expert correlation: r = .67, p < .01, n = 13).

Both novices and experts in the stacked display condition revisited and fixated longer on results, which seems plausible because results sections contain the most relevant information to evaluate whether the provided hypotheses were confirmed or not. Further, as expected in the previous section, the relatively longer return fixation durations in the distributed display condition were observed majorly in methods and results. Thus distributed display users examined critical information in methods and results sections by making returns but stacked display users did it by encoding well at first.

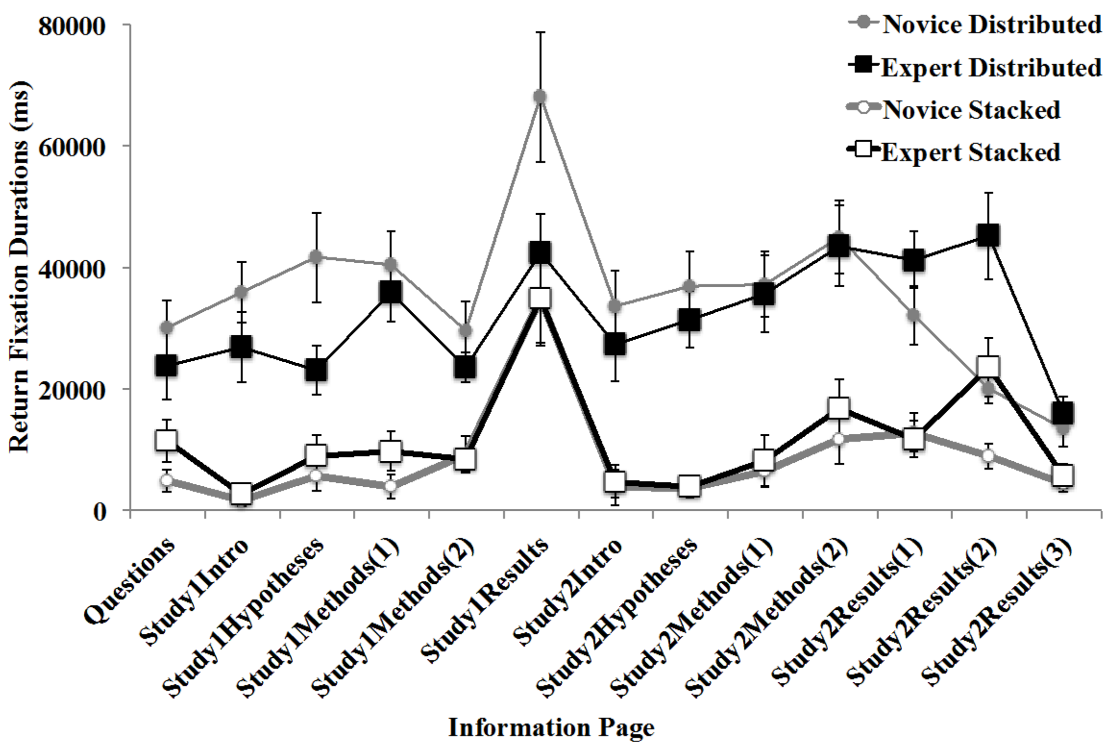


Figure 8. Mean sum of return fixation durations by information page for each condition (with SE bars).

2.2.4 Off Screen Gaze Durations: Externalization of information

Another large possible use of problem solving time involves note taking and note processing. Participants were able to take notes freely during the window examination phase and many of them (although not all of them) took notes and consult these notes while they composed answers. The regular use of notes implies a need for another form of (personally constructed) external storage for this complex task. In addition, since the action of note taking can entail multiple cognitive activities such as selecting information worthy of notes, summarizing, and sometimes writing elaborated ideas or inferences for later use, it can be a place for which expertise may play a particularly important role.

The memorization strategy hypothesis predicts that stacked display users should have spent more time taking notes, essentially as an alternative form of (external) memorization. If people experience relatively higher information access cost when using stacked displays, they may take notes to reduce the need to revisit previous pages and thus to take advantage of cognitive offloading. However, note taking may be governed less by the physical constraints unlike first pass and return fixations because note taking happens at a relatively higher level of cognitive processing than basic information encoding and eye-movement programming. Experts who use distributed displays may make less use of notes because they can selectively and flexibly integrate and compare information across the four quadrants, but novices may feel that it is overwhelming to have information provided across the four spaces at once. Since novices are not skilled enough to manipulate technical and statistical information, too much information may hinder their understanding and the frustration may lead to use of notes as a way to externally reconstruct their own simplified problem space.

To test patterns in note-taking time, off-screen gaze durations longer than 2,000 ms were collected and summed up as a proxy measure of the time spent on note taking (i.e., sum of the time spent on various activities beyond online visual information processing). In general, off-screen gaze durations longer than the normal blink duration of 300-400 ms could have been used as an index of time users are engaged in other activities—which was most commonly note taking in this study. But a more conservative threshold of 2,000 ms was used to reduce the effect of activities other than note taking (e.g. looking at a computer clock to check the time, looks to the mouse after a hand was taken off of them, or head scratches). I use the term note taking to refer to both the writing of new notes, and the mental processing of existing notes.

A 2x2 ANOVA was used to test the effect of display format and expertise level on the sum of off-screen gaze durations longer than 2,000 milliseconds. Detailed statistics are presented in Appendix L. There was no main effect of expertise level, F (1, 95) = 1.20, MSE = 16.47, p = .28, or display format, F (1, 95) = 0.05, MSE = 0.63, p = .83 on the sum of off-screen gaze durations. Interestingly, there was a significant interaction effect, F (1, 95) = 12.59, MSE = 173.38, $p < .001$, $\eta^2 = 0.12$ (see Figure 9). Experts who used the stacked display spent twice as much time on note taking as those who used the distributed display (t (36) = 2.12, $p < .05$, Cohen's d = 0.76), as the memorization hypothesis predicted. But novices showed a reverse pattern: novices who used the stacked display spent less one third the time on note taking than did those who used the distributed display (t (59) = -2.95, $p < .01$, Cohen's d = 0.82).

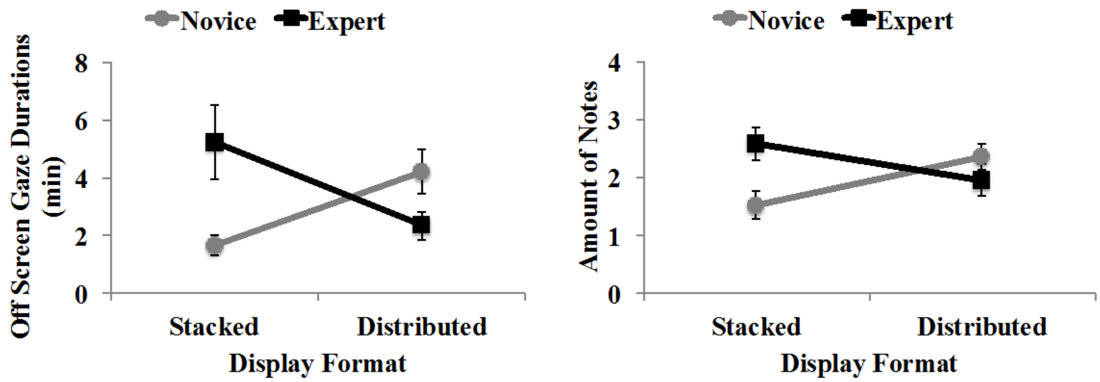


Figure 9. Mean (with SE bars) off screen gaze durations (left) and amount of notes (right) by display format and expertise level.

One may wonder whether it is plausible to consider the off-screen gaze durations longer than 2,000 ms a reliable estimate for the amount of notes taken. To test the idea, the amount of notes for each participant was coded on a four-point scale: little (only a couple of lines), light (5-10 lines), medium (10-20 lines), and heavy notes. The Pearson correlation at the participant level between the total off-gaze durations and the amount of notes was r = .70, n = 99, p < .001. Further, the 2x2 ANOVA on the amount of notes showed exactly the same pattern that was observed in the off-screen gaze duration (see Figure 9). There was neither an effect of expertise, F (1, 95) = 1.64, MSE = 2.59, p = .20, nor an effect of display format, F (1, 95) = 0.14, MSE = 0.22, p = .71, but the interaction was significant, F (1, 95) = 7.87, MSE = 12.43, p < .01, η^2 = 0.08.

The results are consistent with the prediction that note taking time is a measure more heavily influenced by expertise than the two previous eye-tracking-based measures. It seems that experts in the stacked display took more notes as it was relatively harder to access information and they used notes to store necessary information in a readily available place, as the memorization strategy hypothesis would predict. However, novices in the distributed condition took more notes even when multiple information pages were available at once. It may imply that the difficulties novices experienced with a potentially overwhelming amount of information and the lack of expertise might have overridden the effect of the display format.

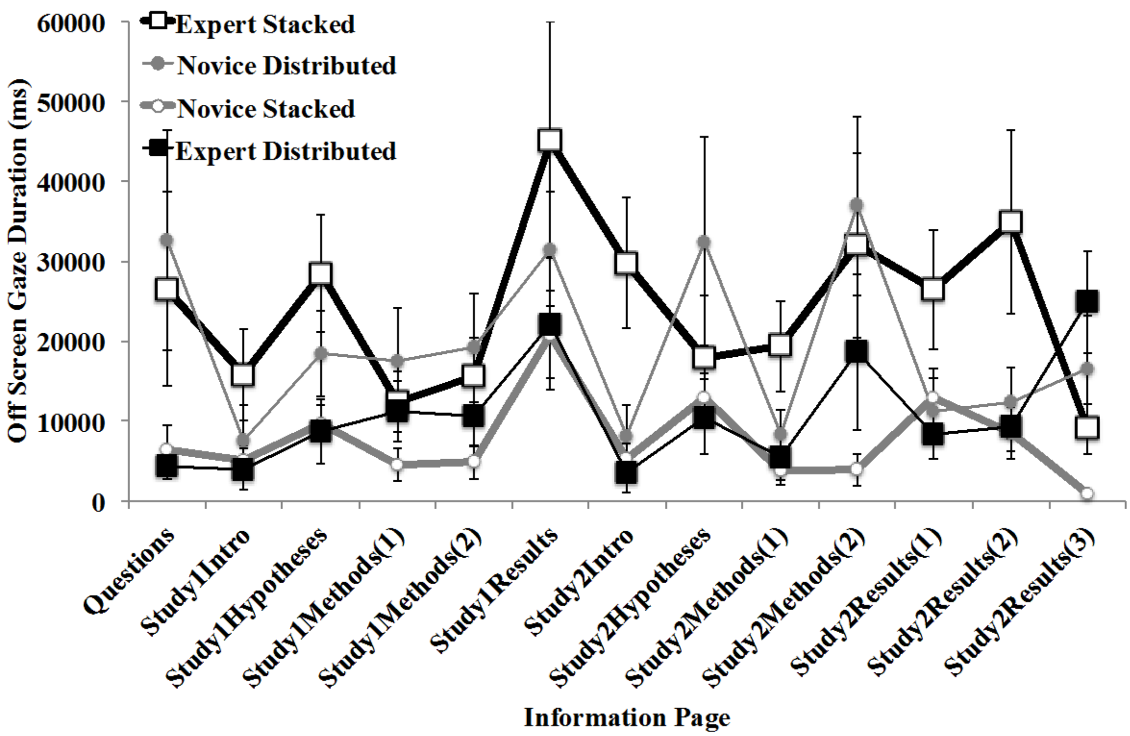


Figure 10. Mean off-screen gaze durations by information page in each condition.

To investigate further on which page problem solvers spent more time engaging with notes, off-screen gaze durations were analyzed by information page (see Figure 10). Overall, participants seemed to spend more time on note taking while processing methods and results pages. Also, as the interaction effect between the display format and the expertise level predicts, experts took more notes when they were using the stacked display and the trend seems consistent across the pages, but novices took more notes when they were using the distributed display and the effect shows a couple of peaks on Questions, Study2 Hypotheses, and Study2 Methods (2). It is rather odd to observe these peaks but it could have been caused by differences in expertise, because it might have been harder for novices to understand hypotheses, methods, and even the questions.

The correlations of the profile across conditions showed a medium stacked-stacked correlation (r = .66, p < .01, n = 13) and a weak distributed-distributed correlation (r = .46, p = .11, n = 13). It appears that for both novices and experts, the effect was driven more by the display format than expertise (very weak novice-novice correlation: r = .33, p = .27, n = 13 and expert-expert correlation: r = .10, p = .76, n = 13).

In addition, it seemed that the time effect mainly came from the quantity of notes rather than the quality. Two qualitative measures were analyzed to see if experts had written more integrative and goal-related notes. As for an integrative note taking, the number of connections made among information from different pages was counted. When contents from different pages were explicitly connected by arrows or written in close proximity, each instance was counted. In general, not so many connections were made. On average, novices made 0.17 (SD = 0.50) connections and experts made 0.53 (SD = 0.72) connections but there was a significant main effect of expertise (F (1, 76) = 7.42, MSE = 2.53, p < .01, η^2 = 0.09). Experts seem to take a bit more integrative notes but the frequency was fairly low. Interestingly, there was a tendency that novices who used distributed display took more integrative notes (distributed: M = 0.22, SD = 0.58, stacked: M = 0.10, SD = 0.30) while experts who used stacked display took more integrative note (distributed: M = 0.40, SD = 0.63, stacked: M = 0.65, SD = 0.79). However, the interaction effect was not significant, F (1, 76) = 1.95, MSE = 0.67, p = .17, η^2 = 0.03. To measure goal-related note taking, the number of written inferences and conclusions was counted. Inferences and conclusions defined as comments that deliver problem solvers' idea, opinion, analyses, and attempt answers for the task questions. There was no effect either by expertise (F (1, 76) = 1.00, MSE = 3.81, p = .32) or display format (F (1, 76) = 0.76, MSE = 2.91, p = .39). Also, there was no interaction effect, F (1, 76) = 1.14, MSE = 4.33, p = .29. Detailed statistics are presented in Appendix M.

2.2.5 Self-Reported Strategies and Cognitive Loads

Strategy and cognitive load survey results showed several patterns that supports previous findings. Detailed statistics are presented in Appendix N. First, novices reported relatively lower expertise on the data interpretation task. In the strategy survey (see Figure 11), regardless of display format they used, novices felt more lost about where they were in the task (M = 1.8, SD = 1.1) than experts reported (M = 1.4, SD = 0.9), F (1, 98) = 4.09, MSE = 4.32, p < .05, η^2 = 0.04, which suggests their unfamiliarity to the data interpretation task and low expertise. It was also found that the invested mental effort (see Figure 12) in understanding theory was larger for novices than experts, F (1, 98) = 6.76, MSE = 17.34, p < .01, η^2 = 0.07 and the perceived difficulty in understanding the results was higher for novices than experts, F (1, 98) = 8.17, MSE = 23.34, p < .01, η^2 = 0.08. More specifically, novices reported lack of skills in integrating information. They not only reported higher perceived difficulty in integrating methods and results, F (1, 98) = 6.65, MSE = 18.91, p < .01, η^2 = 0.06, but also reported that they invested more mental effort to integrate methods and results, F (1, 98) = 9.48, MSE = 28.84, p < .001, η^2 = 0.09.

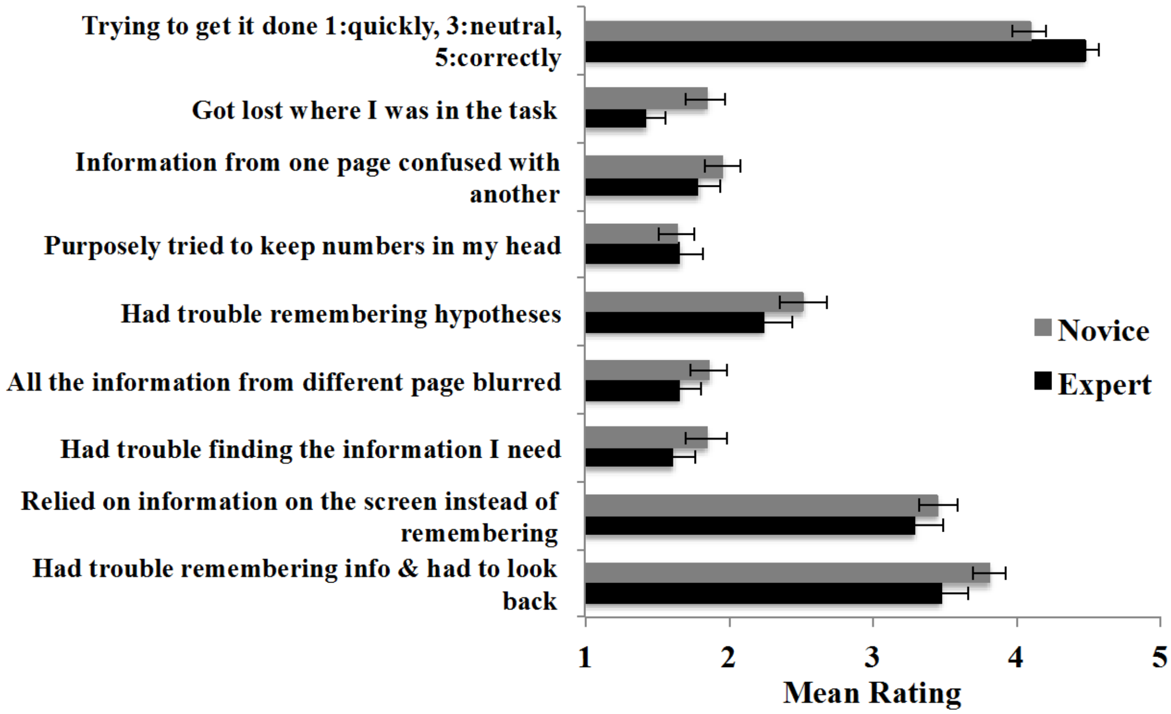


Figure 11. Mean rating on the strategy survey by expertise

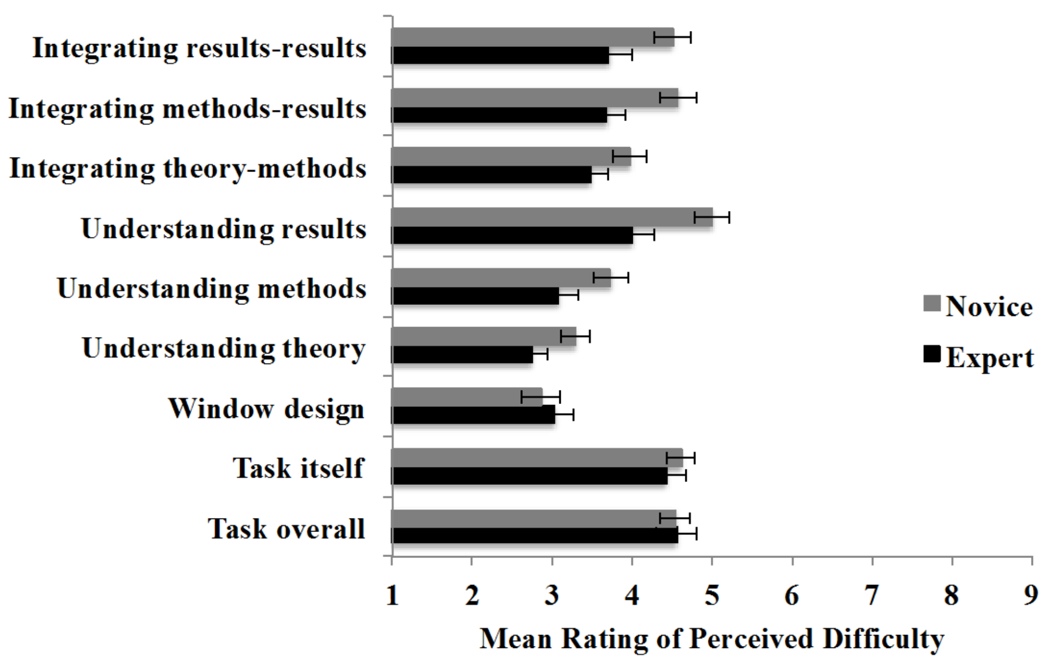


Figure 12. Mean rating of perceived difficulty by expertise

Second, both experts and novices reported difficulty using the stacked display window, which supports the idea that people experience relatively higher information access cost in the stacked display, regardless of expertise. In the strategy survey (see Figure 13), for a question asking if they felt that all the information from different page blurred, stacked display users in both expertise group agreed significantly more than distributed display users did, F (1, 98) = 4.51, MSE = 4.24, $p < .05$, $\eta^2 = 0.04$, which may imply information overload in the head as a side effect of using memorization strategy. In the cognitive load survey (see Figure 14), both experts and novices who used the stacked display reported higher perceived difficulty in using the window design than those who used the distributed display, F (1, 98) = 23.11, MSE = 61.56, $p <$ .001, $\eta^2 = 0.19$, and reported that they invested more mental effort in using the stacked display, F (1, 98) = 18.22, MSE = 49.96, $p < .001$, $\eta^2 = 0.16$.

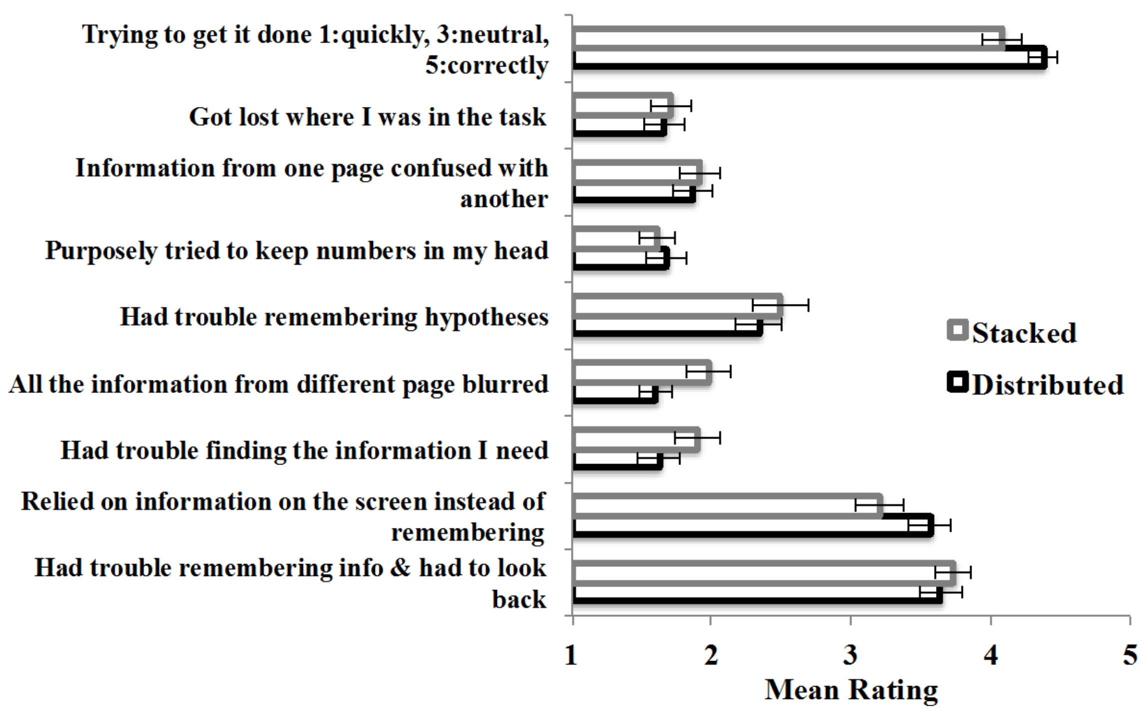


Figure 13. Mean rating on the strategy survey by display format

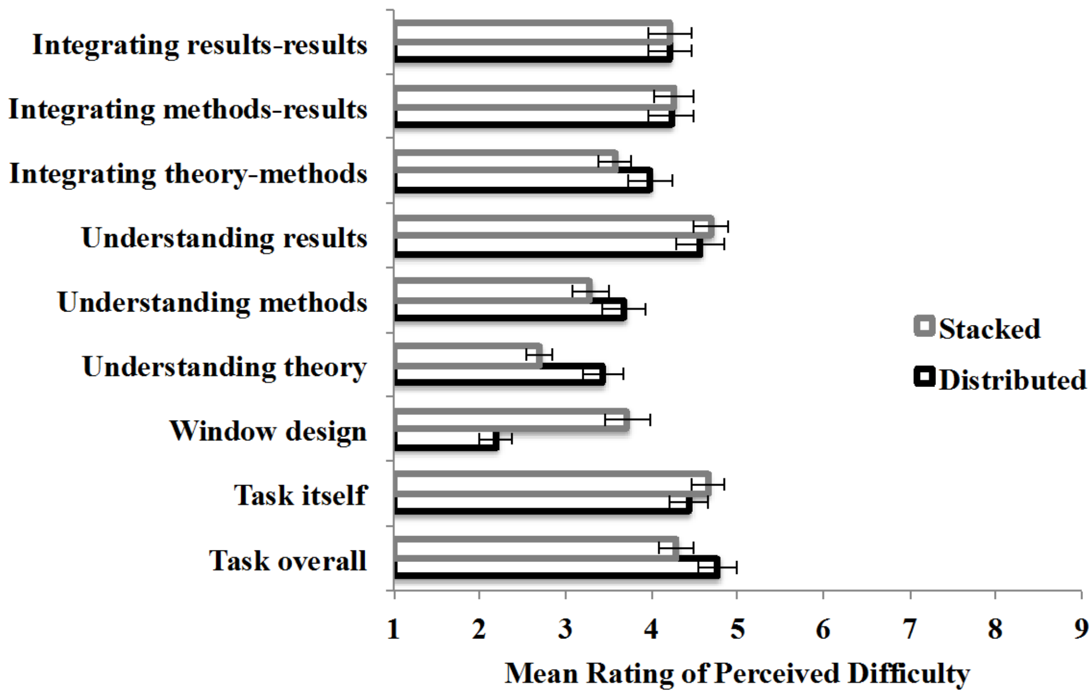


Figure 14. Mean rating of perceived difficulty by display format

Finally, some measures showed interaction effects between display format and expertise that matched to the findings in the previous sections (see Figure 15). In the strategy survey, for a question asking to what degree they had trouble remembering information and had to look back, novices agreed with the statement more when they were using the distributed display, but experts agreed more when using the stacked display, F (1, 98) = 4.12, MSE = 4.23, $p < .05$, $\eta^2 = 0.04$. Further, in the cognitive load survey, novices in the distributed display condition perceived the task more difficult than novices in the stacked display condition while experts perceived the task easier when the task was provided in the distributed display, F (1, 98) = 7.90, MSE = 16.35, $p < .01$, $\eta^2 = 0.08$. And the same pattern was observed in the reported amount of mental effort they invested to solve the task, F (1, 98) = 14.91, MSE = 29.92, $p < .001$, $\eta^2 = 0.13$.

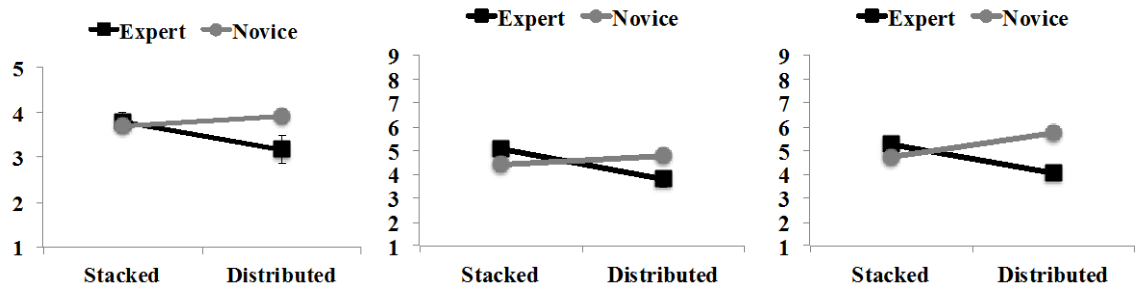


Figure 15. Mean ratings by display format and expertise for a strategy survey question (left) asking how much people agree that they had trouble remembering information and had to look back, perceived difficulty of task itself (middle), and mental effort invested to solve the task (right).

2.2.6 Integrating the Effects of the Three Main Time Elements

The difference observed in the total window-examining time can be explained by the interplay of effects on the three different main times examined thus far: time spent on first pass fixations, time spent on return fixations, and time spent on note taking. There were no correlations above 0.14 between the three factors.

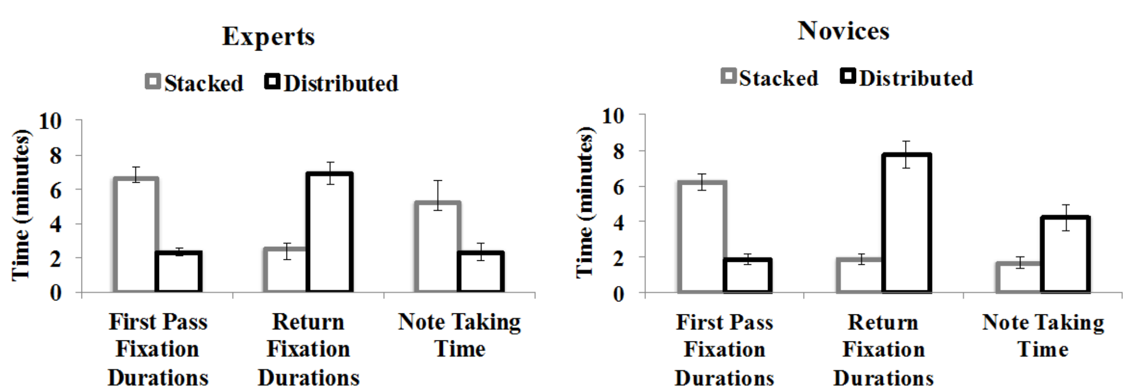


Figure 16. First pass and return fixation durations and note taking time by display format for experts (left) and for novices (right)

First, as the memorization hypothesis predicted, problem solvers spent more time on first time encoding and less time on return fixations with stacked displays, and this pattern was found true regardless of expertise (see Figure 16). That is, the stacked display produced more information internalization due to its higher information access cost. Note that a similar pattern could be observed when analyzing the number of first pass and return fixations, rather than fixation durations. Across both levels of expertise, stacked display users made more first pass fixations and fewer return fixations while distributed display users showed exactly the reverse pattern (number of first pass fixations: F (1, 95) = 128.57, MSE = 22,835,437, $p < .001$, $\eta^2 = 0.58$ and number of return fixations: F (1, 95) = 85.94, MSE = 32,952,317, $p < .001$, $\eta^2 = 0.48$). Similarly, analyses of page transition frequency (i.e., a fixation-based measure including page-turning action: how many times participants moved their eyes to different pages) showed that both experts and novices in the stacked display condition made far fewer page transitions (M = 22.3, SD = 6.2) than those in the distributed display condition (M = 153.3 SD = 70.6), F (1, 95) = 148.75, MSE = 391,427, $p < .001$, $\eta^2 = 0.61$. Detailed statistics are presented in Appendix O.

Second, note taking was found to be interactively affected by display format and expertise. Unlike the first two factors that were directly governed by the physical constraint and thus showed the same pattern of effect across the two expertise level, the analyses of note taking time and amount showed that experts took more notes when they were using the stacked display but novices took more notes when using the distributed display. Given the knowledge and skills that experts have, it could have been beneficial for experts to have four pages of information laid side by side so that they can compare and integrate information across hypotheses, methods, and results. The stacked display, in contrast, might have hindered experts because the limited availability of information could not support the fast flow of thoughts. In a reaction to avoid discrete information availability in the stacked display, experts seemed to take more notes so as to externalize the necessary information in a handy place. However, due to the lack of knowledge and experience, novices might have felt overwhelmed by having too much information available in the distributed display, not knowing where to focus. The larger amount of note taking observed in the distributed condition may be the result of this frustration and an effort to reorganize their thoughts by copying the information in a new format. Thus, it could have been actually easier for novices to have each information page displayed separately and digested in a more step-by-step manner.

Finally, taken altogether, the overall difference observed in the window examining time can be explained by the combined contribution of the three factors (see Table 6). The time difference between stacked and distributed displays observed in each factor was computed and summed up, then compared to the time differences observed in the window time. For novices, the mean window time difference was 13.6 - 19.2 = -5.6 minutes, and the three factors explained 70% of this time difference. For experts, the mean window time difference was 20.1 - 15.7 = 4.4 minutes, and 64% of this time difference was explained by the three factors.

Table 6. Mean time differences in first pass fixation durations, return fixation durations, and note taking time (in seconds). Sum of the effect of the three factors is computed for novices and experts.

	Novice			Expert		
	Stacked	Distributed	Stacked – Distributed	Stacked	Distributed	Stacked – Distributed
First pass fixations	6.2	1.8	4.4	6.6	2.3	4.3
Return fixations	1.9	7.7	-5.8	2.5	6.9	-4.4
Note taking	1.7	4.2	-2.5	5.2	2.3	2.9
Sum	9.8	13.7	-3.9	14.3	11.5	2.8

About a minute and half was left unexplained in both groups. Given that the note taking time computed only by off screen gaze durations longer than 2,000 ms, the unexplained time may include time spent on multi-tasking other than note taking such as information tracking and even some eye blinking when people make fixation jumps within/between pages. In addition, information-loading time (i.e., time required to load and view an information page whenever a selection is made) was not accounted for the equation.

The number of physical page transitions (i.e., page-turning action only: how many times participants selected and changed information page to view using the drop down menu) showed that both experts and novices made more page transitions in the stacked display condition (M = 22.8, SD = 6.5) than when they were in the distributed display condition (M = 19.7, SD = 6.4), F (1, 98) = 8.14, MSE = 323.02, p < .01, η^2 = 0.08 and this tendency was stronger for experts (Stacked: M = 25.7, SD = 6.6 vs. Distributed: M = 19.4, SD = 5.7) than novices (Stacked: M = 20.9, SD = 5.8 vs. Distributed: M = 19.9, SD = 6.9), F (1, 98) = 4.16, MSE = 164.98, p < .05, η^2 = 0.04. Presumably, people in the distributed display condition did not have to turn pages so frequently once they made available the four information pages they needed. They could simply move their eyes to access information presented in the four quadrants. People in the stacked condition, however, had to turn pages if they wanted to look at information in another page and they seemingly turned more pages than people in the distributed display condition, as the task requires deep thinking and integration. But they were reluctant enough not to turn the pages four times more than the number of page turned in the distributed display condition; in fact, much less.

The relatively stronger tendency in experts may imply that experts were more prudent and made more page transitions to check information even if it was bothersome in the stacked display condition. More interestingly, experts in the stacked display condition might have made more page transitions because they wanted to *integrate* information across pages and the only way that they can achieve the goal was to go back and forth even though it was cumbersome; likewise, the physical page transition time can explain a portion of the unexplained time for experts. For novices, whether having a distributed display or not may not matter much (or even worse to have a distributed display) since they are not sufficiently trained to readily make connections across information in different sections. But for experts, having a distributed display does matter because it helps them cross-examine and integrate information across pages, which is the most crucial part of the successful problem solving for a data interpretation task.

3.0 GENERAL DISCUSSION

3.1 DISTRIBUTED VS. STACKED DISPLAYS

Replicating a numbers of prior studies with artificial tasks (Jang & Schunn, 2012; Jang et al., 2011) and a couple of studies with authentic tasks (Jang et al., 2012; Kroft & Wickens, 2002), the current work again finds that the organization dimension of distributed versus stacked displays can have a large effect on problem solving performance. The current study carefully controlled display content to manipulate display format *per se*, now establishing that even with careful controls, the dimension matters even with rich, diverse tasks and with participants who are well trained in the task. Thus, I have increased confidence that this dimension is important in applied settings. As a result, the trends towards using micro-displays for work applications (powerful ultra-light laptops and tablet computers) should be re-examined carefully with respect to impact on performance. Further, it is recommended for optimizing software to better enable information distribution on the screen.

3.2 EFFECT REVERSAL AND EXPERTISE (DOUBLE) REVERSAL

Prior research on the distributed vs. integrated spatial arrangement effect found an expertise reversals such that experts actually benefit more from what would be worse for novices

(Kalyuga, Ayres, Chandler, & Sweller, 2003). Studies on other factors impacting task performance (Cho, 2004; Yeung et al., 1998) also showed what is the best for experts may not do the same for novices. Therefore, a primary goal of the current study was to examine whether the benefit of distributed over stacked displays held for experts. Regarding this goal, the study did reveal a robust interaction between spatial arrangement and expertise. However, details of this expertise interaction were superficially quite surprising. Compared to previous studies (Jang & Schunn, 2012; Jang et al., 2011; Jang et al., 2012) that showed consistent task completion time benefits of distributed displays with relatively untrained participants, the current study actually found the opposite overall task completion time effect for novices. And then the display effect on task completion time for experts in this study did reverse what was found for novices on the same task, but back to what was previously found with novices on other tasks.

The opposite task time effect can be explained by the differences in tasks. In previous lab tasks, individuals were able to be experts for the given tasks with a brief practice before the main task began because there was a relatively simple set of procedures and well-taught integration strategies to follow. Unlike previous lab tasks, this time novices working on an ill-defined problem that was quite difficult given their domain-specific and domain-general knowledge. Presumably, novices might have experienced two types of difficulty such that they do not know how to digest the content itself and how to integrate information across pages, if they even realized that integration is crucial to the problem solving.

Further, the current stacked better than distributed effect on overall time with novices working scientific data analysis tasks does have some prior basis, albeit in preference rather than performance data. In a previous study of student format preferences (Jang & Schunn, 2011), psychology lab undergraduates were trained on statistical analysis procedures for a couple of weeks and then asked to choose and use either distributed or stacked format of statistic instructions to analyze a data on their own. About 70% of students chose the stacked format of instruction, which was a surprising overall ratio given that the distributed format had previously produced more efficient problem solving and greater learning outcomes (Jang et al., 2011). However, the current study suggests that the preference for stacked format of instructions among undergraduates may have been an adaptive choice for their level of expertise. Even though students were trained on the domain knowledge for a few weeks, the time and experience might have been short for them to confidently choose and use distributed information.

I argue that the reversed novice pattern and then double reversal for experts found in the current study is only superficially surprising. Using the theoretical decomposition of factors influenced by distributed versus stacked displays that was introduced in this paper, it is possible to show how the pattern of effects across past and current results can be understood. The eye-tracking data provides more direct access to the multiple factors that are in play, and this new level of theoretical precision and methodology for tracking underlying factors should enable more robust theorizing and application in various other domains. In the next section, I unpack aspects of expertise that need to be considered in our theoretical account, and then I discuss the larger underlying theory in the section after that.

3.3 UNPACKING EXPERTISE EFFECTS

Several expertise characteristics are relevant to the current study. It is important to note that graduate students participated in the current study are perhaps best called near-experts. Although the psychology graduates had 5 years of experience on average and most of them had firstauthored journal publications, which clearly distinguishes them from novices, they did not meet the commonly used 10-year rule (Hayes, 1985) definition of world-class expertise. But expertise is really a continuous dimension that just a binary expert/novice distinction (Schunn, Saner, Kirschenbaum, Trafton, & Littleton, 2007), and for practical applications, it is important to understand effects along the broader continuum, especially since more people sit in the middle than at the end points. At the same time, the literature on expertise has focused on end points, and here I discuss two relevant expert behaviors, specifically regarding information search strategy and information representation.

First, experts selectively access relevant information. Expertise involves learning and developing an eye for information value and probability: where to look and where not to look (Chi, Feltovich, & Glaser, 1981; Hinsley, Hayes, & Simon, 1977; Patel & Groen, 1991; Spilich, Vesonder, Chiesi, & Voss, 1979). With the same basic limited cognitive capacity shared across all humans, experts become more efficient by attending to and consuming the types of information that provide high information value and are essential to problem solving. The graduate students in our study appeared better able than the undergraduates to manage and selectively view the information they needed to integrate among the multiple sources of information that consisted of graphs, tables and texts. Further, the trained graduate student eyes presumably focused less on tangential information within and across pages, thereby reducing cognitive load.

Second, experts are often found to have functional, abstracted representations of presented information (Moss, Kotovsky, & Cagan, 2006). For example, expert physicists grouped physics problems according to which principles and equations are useful for solution, whereas novices grouped problems according to similarities in superficial features (Chi et al., 1981). Similarly, when asked to recall half-innings from a baseball game, ardent baseball fans structured the game by major goal-related sequences of the game, such as advancing runners, scoring runs, and preventing scoring, rather than the less essential components (e.g., weather or crowd mood) that novices often used in recall (Spilich et al., 1979). Based on these observations, the graduate students in the current study should have benefited from building a structured representation of the data interpretation problem, which in the current study can be indexed by note-taking behaviors–"task reflection as participants attend to problems" (Chi, 2006, p. 176). The content of the graduate students' notes showed a tendency to be more integrative and goal-related (e.g., writing key information, ideas instead of verbatim, new inferences, and connections among information from different pages with arrows or by writing them in close proximity), compared to a sequential list of exactly copied information that was shown often in the notes of novices. However, the expertise effect on note content was not large presumably because graduate students were not very far down the expertise continuum from the undergraduates.

It may seem odd that the graduate students did not appear to benefit in this task from the well documented chunking benefits associated with expertise (Chase & Simon, 1973). However, it is important to note the current task was not only integrative but also relatively ill defined. Although the graduate students were trained on relevant analytic scenarios and schemata, ill-defined tasks such as data interpretation of scientific studies in a new domain will have few familiar chunks to re-use. It is likely that analysis of data in their own focal areas of expertise would have presented the opportunity to use familiar chunks (Schunn & Anderson, 1999).

3.4 UNPACKING THE UNDERLYING MECHANISMS OF DISPLAY ORGANIZATION EFFECTS

The overall task time effects of display organization needs to be understood as the summation of three different underlying factors that influence task time: information internalization time, information access time, and information externalization time (see Figure 3).

Stacked displays lead to higher information internalization time. This effect is indexed by mean fixation times during first pass through content, and it appears to hold for both novices and experts. Other researchers (Gray & Fu, 2004; Gray, Sims, Fu, & Schoelles, 2006) have conceptualized this effect as a memorization micro-strategy that is automatically applied when there are relatively high external information access costs, like in the case of stacked displays.

At the same time, in complementary fashion, stacked displays actually lead to lower information access times for both novices and experts, as indexed by shorter total return fixations. As a result of not having memorized the information, problem solvers using the distributed display must then more often search external for information to integrate, whereas the problem solvers using the stacked displays can rely on internally stored information, which will typically (although not always) be accessed more quickly than external information. It is important to note however, that the information internalization effect is larger than the information access effect, and thus, on the basis of just these two effects, there should generally be an overall time benefit of distributed displays over stacked displays. However, in the case of particularly complex integrative tasks that require constant revisiting, it could be that the stacked displays will be more beneficial because of the high frequency of revisiting previously memorized (or not) information.

The third effect, informational externalization time appears to be modulated by expertise level, but this is likely expertise level relative to the complexity of the task. In very simple tasks (such as what was studied previously), or for more expert problem solvers in more complex tasks, problem solvers take more notes with stacked displays, probably as an effort to manually make the information distributed and reduce the need to re-access information. In some ways, this effect could also be seen as an adaptive tradeoff between externalization time and information access time. By contrast, novices given a complex task took more notes in the distributed displays condition. These additional notes may have been taken to compensate for the cognitive overload from being presented so much difficult to integrate information: the additional notes constitute an attempt at creating a simplified information space. Novices in the distributed display condition did report higher cognitive load (perceived difficulty and mental effort). Alternatively, the novices may have taken fewer notes in the stacked display case because they saw less information that needed to be integrated; that is, they were less aware of the complexity of the integration task at hand. Indeed, other research has found that undergraduates often fail to even recognize the need for experimental data to be integrated with an overarching theory (Schunn & Anderson, 1999, 2001).

3.5 IMPLICATIONS

The research has several implications for the field of cognitive psychology, human factors, and education. This study examined an ill-defined task, which is much more common in the work place than studied by researchers. Further, it specifically examines a kind of task that is common across most domains of science. That is, the basic skill sets required for accurate quantitative data interpretation are common to a range of social sciences and human science fields, such as medicine, public health, marketing, economics, political science, and sociology. Further, there are important commonalities to quantitative data analysis in the natural sciences and engineering. Also, this study is relatively rare in directly comparing (more) expert and (more) novice problem solvers, and it showed a clear expertise reversal effect that is important to consider in proscribing implications for visual display use and design, and for the development of educational materials and strategies. For more expert problem solvers, the overall trend toward using smaller displays appears to be counterproductive to problem solving performance, whereas for more novice learners working on complex material, their use of laptops and tablets that naturally limit the amount of information displayed may be adaptive.

In terms of theoretical implications, the study generated an understanding of the complex relationships between cognitive load, reasoning processes, and problem-solving performance from students to professionals. The proposed tripartite framework regarding underlying mechanisms was found to be useful for explain the time differences and it is expected to be applicable to further unpack the effects of other display type comparisons (e.g., stacked vs. integrated displays or collaborative information sharing systems). The underlying framework includes factors at the lower, perceptual level, but also high-level cognitive processes, and the overall model considers the bi-directional interaction between the information presentation environment (e.g., innate high information access cost in stacked displays) and the human's activities (e.g., strategic information encoding and note taking).

The proposed research also leads to practical suggestions for improvements to statistical analysis and data visualization software tools. For example, data output displays typically used by scientists can be improved in a way that better supports users' data comprehension and interpretation. Many students and colleagues are dissatisfied with the data visualization of statistical analysis software, which tend to show tables and graphs in a sequential and stacked manner. Even when people try to print out data and set up their own versions of a distributed display (e.g., spread the pages out on a table), often the tables and graphs do not print correctly or take too many pages to be printed. Considering the large number of users of data analysis software, including students, businessmen, scientists, and policymakers, data visualizations that better facilitates comprehension and problem solving could have a broad overall impact.

3.6 FUTURE DIRECTIONS

This research suggests that spatially distributed information lead experts to more efficient and effective problem solving. For future studies, boundary conditions of the effect (e.g., expertise and distribution ratio) should be explored to deepen our understanding of cognitive mechanisms and to provide precise recommendations for designers. For example, is there a benefit of distributing information across two large monitors rather than just two large monitors? Or does the distributed benefit hold for 20-year experts working on data in their own focal area of expertise? At the other end of the expertise continuum, given that novices in this study could not benefit from distributed displays, a question is remaining on how much expertise is required to benefit from distributed displays. Relevant subquestions include whether it requires domain-specific, domain-general knowledge or a combination of both and how individual differences in cognitive abilities and strategy adaptivity would affect the information encoding strategy. As Schunn and Reder (2001) found, some people are much less sensitive and slow to change their strategy choices.

APPENDIX A

TOPIC PASSAGE FOR PRACTICE

PLEASE READ OUT LOUD ONCE, AND READ IT OVER UNTIL YOU CAN PARAPHRASE THE HYPOTHESIS IN YOUR OWN WORDS BECAUSE YOU WILL BE ASKED TO DO SO

Effect of Diagram on Learning

There is a Chinese saying that a picture is worth 10,000 words. Like many bits of common wisdom, it turns out this is sometimes true. The task for cognitive psychologists is to figure when it is true and when it is not true.

A lot of that previous work has been done on the effect of having pictures and diagrams and found significant benefits. For example, college students were found to learn better when a science text is accompanied with a relevant diagram.

Willows (2005) recently conducted two studies to examine the effect of diagram on children's learning (i.e., information recall and speed and accuracy of reading).

Hypotheses

H1: Pictures would facilitate recall of information.

H2: Pictures would improve the speed and accuracy of reading.

YOU WILL BE PROVIDED WITH WILLOWS'S HYPOTHESES, METHODS, AND RESULTS. YOUR TASK IS TO INTERPRET THE RESULTS IN ORDER TO ANSWER GIVEN QUESTIONS.

APPENDIX B

MATERIALS FOR PRACTICE

Note: For readers' convenience, page titles are placed at the top in bold and the end of each information page marked as "*(Page Ends)*" in bold. In the experiment, the titles were shown in a drop-down menu and "*(Page Ends)*" were not shown for participants.

Questions

You will be asked to type your answers for these questions once you said you are done interpreting the results and ready to give your answers. You can refer back to your notes and task window while answering.

1. Was the hypothesis of Study 1 confirmed? If so, what are the evidence? If not, what are the evidence?
2. Was the hypothesis of Study 2 confirmed? If so, what are the evidence? If not, what are the evidence?
3. Are results of Study 1 and 2 congruent? If so, in what ways and if not, in what ways?
4. If you've answered that there was an inconsistency in the findings of the two studies, How would you reconcile the findings?

(Page Ends)

Study 1: Intro & Hypothesis

The purpose of Study 1 was to investigate the effect of pictures on the recall of expository prose by 1st graders.

H1: Pictures would facilitate recall of information.

(Page Ends)

Study 1: Methods

Thirty-three first graders were presented with information about unusual animals in one of two conditions.

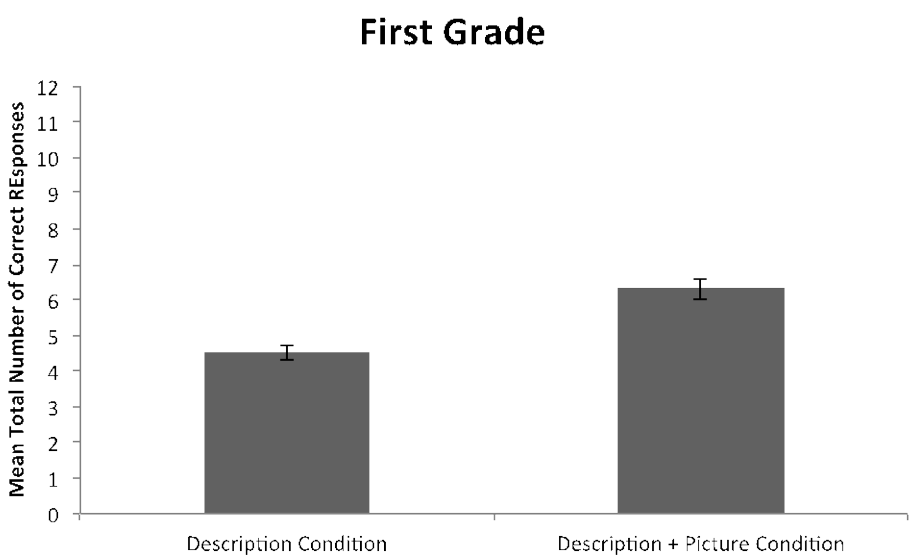
Description condition: children listened to descriptions of the animals

Description plus picture condition: children listened to a description while they looked at a picture of the animal

Dependent variable: Cued recall test of information

(Page Ends)

Study 1: Results



(Page Ends)

Study 2: Intro & Hypothesis

On the other hand, Opinions among educators and authors of children's beginning readers about the value of illustrations range from those who believe that pictures serve an essential function in the instructional process to those who believe that pictures serve no useful purpose and that they may interfere with children's learning to read (Aukerman, 1971; Chall, 1967).

H2: Pictures would improve the speed and accuracy of reading.

(Page Ends)

Study 2: Methods

Thirty-two second graders were asked to read aloud a list of words by with either no pictures or associated pictures.

In the control condition, the words were simply printed on the page in three rows of five.

In the picture condition each of the 15 words was superimposed on a related picture, for example the word cat was superimposed on an outline of a dog. Subjects were asked to read the words on the page aloud as quickly as possible.

Dependent variables: time and accuracy (# of errors)

(Page Ends)

Study 2: Results

Variable	No pictures	Pictures
Time (sec)		
M	86.59	103.81
SD	53.76	61.82
Errors		
M	6.53	10.56
SD	6.03	5.38

(Page Ends)

APPENDIX C

TOPIC PASSAGE FOR TASK

PLEASE READ OUT LOUD ONCE, AND READ IT OVER UNTIL YOU CAN PARAPHRASE THE HYPOTHESIS IN YOUR OWN WORDS BECAUSE YOU WILL BE ASKED TO DO SO

Destination Memory vs. Source Memory

Everyone has recounted a story or joke to someone only to experience a nagging feeling that they may already have told this person this information. It is for this reason that people sometimes preface a story with ‘‘stop me if I’ve told you this before.’’ Remembering to whom one has told things not only can help one avoid social embarrassment, but also may be critical in some situations. For example, supervisors need to remember to whom they told specific information or delegated particular responsibilities so that they may assess progress and accurately gauge employees’ workloads, and liars need to keep track of the information that they have told to particular people to avoid getting caught telling incongruent stories. Remembering to whom one has told things also is necessary for facilitating everyday interactions, such as conversations with friends. People can assume a common ground and continue where they left off only if they remember what they told to different friends (cf. the given-new contract—Haviland & Clark, 1974). Consequently, in daily interactions, people need to remember not only who told them things, or the source of information, but also to whom they told things, or the destination of information.

The processes involved in remembering the source of information (e.g., in conversations, who told you something) have been comprehensively studied and are referred to as source memory (Johnson, Hashtroudi, & Lindsay, 1993; for a review, see, e.g., Mitchell & Johnson, 2000). Studying source memory makes sense, given the importance attached to remembering sources. For example, remembering that information was obtained from CNN rather than MTV is likely to determine how that information is used. Yet the inverse situation—remembering the people one has told something to—is often important as well. Thus, it is surprising that researchers know very little about the processes involved in remembering the destination of information that people output. We refer to these processes, by analogy, as destination memory.

Gopie (2009) recently conducted two studies to examine which memory is more fallible and why.

Hypotheses

H1: Source memory is more fallible because self-generated information is usually better remembered

H2: Destination memory is more fallible because outgoing information is not as well integrated with its context (i.e., the person whom one tells a fact)

YOU WILL BE PROVIDED WITH GOPIE'S HYPOTHESES, METHODS, AND RESULTS. YOUR TASK IS TO INTERPRET THE RESULTS IN ORDER TO ANSWER GIVEN QUESTIONS.

APPENDIX D

MATERIALS FOR TASK

Questions

You will be asked to type your answers for these questions once you said you are done interpreting the results and ready to give your answers. You can refer back to your notes and task window while answering.

1. Was the hypothesis of Study 1 confirmed? If so, what are the evidence? If not, what are the evidence?
2. Was the hypothesis of Study 2 confirmed? If so, what are the evidence? If not, what are the evidence?
3. Are results of Study 1 and 2 congruent? If so, in what ways and if not, in what ways?
4. If you've answered that there was inconsistency in the findings of the two studies, how would you reconcile the findings? In other words, what do you think could account for these inconsistent results? Propose at least one hypothesis about the results.

(Page Ends)

Study 1: Intro

Self-generated information is usually better remembered than other-generated information (Slamecka & Graf, 1978). Therefore, memory to whom something was delivered might also be better than memory from whom it was received, all other factors being equal. At this juncture it remains an open empirical question of whether source memory or target memory differ from one another because they have never been directly compared. In addition, if they do differ, then no existing theory specifies whether memory might favor source versus target information.

Gopie conducted a direct comparison of source memory (i.e., information input) to target memory (i.e., information output) holding all other experimental variables constant. His hypothesis was that input and output are differentially remembered only when a decision component is involved.

(Page Ends)

Study 1: Hypothesis

The purpose of Study 1 was to test conditions in which participants received and gave away equal numbers of objects from two fictitious people. The use of male versus female sources has a long tradition in the source-monitoring literature (e.g. Ferguson et al. , 1992; Johnson et al. , 1995). Consequently, Gopie decided to use fictitious male and female names as the sources and targets in this experiment.

H1: Source memory is more fallible because self-generated information is usually better remembered. Specifically, memory for giving someone an object should be better than memory for receiving an object because giving an object involves a decision.

(Page Ends)

Study 1: Methods (1)

Within-subject design and 18 undergraduates participated.

Source monitoring condition: 60 objects (e.g., book, telephone) were presented with half of the items randomly assigned to each of the two female sources (from Sally or from Mary)

Target monitoring condition: an object label appeared in the center of the screen and participants had been instructed to press a key to give away the object to either of the two male sources (to Derek or to Robby); participants were instructed not to use any special strategy such as assigning all of the objects from one of the females to one of the males ***(Page Ends)***

Study 1: Methods (2)

Memory test: 30 new items were intermingled with the old items as distractors; an object label was presented and the query ‘From Sally or Mary?’ appeared beneath it. Participants pressed one of three keys to indicate from whom they thought they had received the object (or was new). Then the query ‘To Derek or Robby?’ appeared for participants to respond.

Independent Variables: source and target monitoring

Dependent Variables: proportions of correct identifications and confusions

(Page Ends)

Study 2: Results

n = 18	Source monitoring		
Responds	New	Sally	Mary
New	0.8	0.1	0.1
Sally	0.1	0.5	0.4
Mary	0.1	0.5	0.5
n = 18	Target monitoring		
Responds	New	Derek	Robby
New	0.8	0.1	0.1
Derek	0.1	0.7	0.2
Robby	0.1	0.2	0.8

n: number of subjects in the condition

(Page Ends)

Study 2: Intro

On the other hand, previous research demonstrated that encoding of the external environment is disrupted when actions are performed by oneself rather than by someone else. For example, Koriat et al. found that when participants performed, as opposed to watched, such as raising their hands or stirring water in a cup, their memory for the context (i.e., the room in which the task was performed) was worse for self-performed tasks than for other-performed tasks. On this basis, Koriat et al. proposed that output events are not as well integrated with their environmental context as are input events. In the case of incoming information, rich associative links are formed between an event and its environment. In contrast, output events are less strongly integrated with their context because people perceive their own behavior as belonging more to themselves than to their environment. Consequently, for output events, people associate their behavior with their internal mental processes rather than with the environment.

(Page Ends)

Study 2: Hypothesis

The goal of Study 2 was to investigate which memory (destination memory or source memory) is more error prone than source memory. Gopie accomplished this by having participants either tell facts to pictures of famous people (destination memory episodes) or learn facts from pictures of famous people (source memory episodes). Subsequent recognition tests assessed memory for individual components of these episodes and for destination memory or source memory.

H2: Destination memory is more fallible because outgoing information is not as well integrated with its context (i.e., the person to whom one tells a fact) as is incoming information (i.e., the person from whom one learns a fact).

(Page Ends)

Study 2: Methods (1)

Sixty undergraduates participated. Half was randomly assigned to each condition.

Destination condition: participants were instructed to tell the facts to the faces. After a fixation cross, a fact was presented. After reading the fact silently, the participant pressed a key, which resulted in a blank screen, followed by a color picture of a famous person. The participant was to tell the famous person the fact that he or she had just read. This procedure repeated until the participant had told each of the 50 facts to a different face.

Source condition: participants were instructed that facts would be told to them by famous people. After a fixation cross, a famous person's face appeared. After viewing the famous face, a key press resulted in a blank screen, and then a fact. The participant read the fact that the depicted famous person was ‘‘telling’’ him or her. This procedure repeated until the participant was told each of the 50 facts by a different face

(Page Ends)

Study 2: Methods (2)

Item memory test: 20 facts and 20 faces (half of which participants had studied, and half of which they had not studied) were randomly ordered and individually presented. The participant responded *yes or no* whether that item had appeared during the study phase.

Associative memory test: 40 face-fact pairs were shown in random order: Twenty pairs had been presented during the study phase, and the other 20 were random re-pairings. Participants reported whether they had previously told that fact to that face (destination condition) or whether that face had told them that fact (source condition).

Independent Variables: condition (destination or source) and item type (face, fact, and face-fact pair)

Dependent Variables: hits (rate of correct *yes*), false alarms (rate of incorrect *yes*), and correct recognition (proportion of hits minus proportion of false alarms)

(Page Ends)

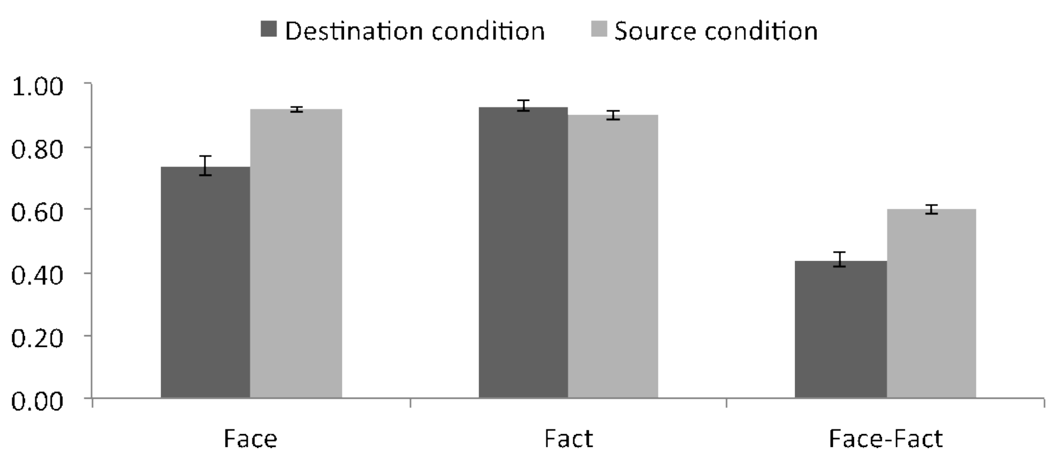
Study 2: Results (1)

Condition and item type	Hits	False alarms (FA)
Destination condition (n = 30)		
Face	.8	.1
Fact	.9	.1
Face-fact pair	.7	.3
Source condition (n = 30)		
Face	.9	.1
Fact	.9	.1
Face-fact pair	.8	.2

Hits: face/fact/pair appeared before and participants said yes
FA: face/fact/pair did not appear before and participants said yes
n: number of subjects in the condition

(Page Ends)

Study 2: Results (2)



(Page Ends)

Study 2: Results (3)

There were significant main effects of study condition, F (1, 58) = 12.02, p = .001, $\eta 2$=.17, and of item type, F (2, 116) = 122.65, p < .001, $\eta 2$ = .68, as well as a significant interaction of study condition with item type, F (2, 116) = 9.77, p < .001, $\eta 2$ = .14.

(Page Ends)

APPENDIX E

TASK QUESTION ANSWER WINDOW

Destination Memory: Task Questions

Exit this survey

*** 1. Name:**

*** 2. Was the hypothesis of Study 1 confirmed?**
If so, what are the evidence? If not, what are the evidence? Be specific.

- Yes, the hypothesis was confirmed
- No, the hypothesis was not confirmed

Your Explanation

*** 3. Was the hypothesis of Study 2 confirmed?**
If so, what are the evidence? If not, what are the evidence? Be specific.

- Yes, the hypothesis was confirmed
- No, the hypothesis was not confirmed

Your Explanation

✱ 4. Are the results of Study 1 and 2 congruent? If so, in what ways and if not, in what ways? Be specific.

- ◯ Yes, the results of Study 1 and 2 are consistent
- ◯ No, the results of Study 1 and 2 are inconsistent

Other (please specify)

✱ 5. If you've answered that there was an inconsistency in the findings of the two studies, how would you reconcile the findings? In other words, what do you think could account for these inconsistent results? Propose at least one hypothesis about the results.

Done

APPENDIX F

STRATEGY SURVEY

Strategy Questionnaire

Exit this survey

***1. Name:**

***2. During problem solving,**
I had trouble remembering information from prior pages and often had to look back.

	Disagree		Neutral		Agree
Agreement	○	○	○	○	○

***3. During problem solving,**
I generally relied on information being available on the screen instead of trying to remember it.

	Disagree		Neutral		Agree
Agreement	○	○	○	○	○

***4. During problem solving,**
I had trouble finding the information I need.

	Disagree		Neutral		Agree
Agreement	○	○	○	○	○

***5. During problem solving,**
All the information from different pages blurred together.

	Disagree		Neutral		Agree
Agreement	○	○	○	○	○

***6. During problem solving,**
I had trouble remembering the hypotheses.

	Disagree		Neutral		Agree
Agreement	○	○	○	○	○

***7. During problem solving,**
I purposely tried to keep numbers in my head.

	Disagree		Neutral		Agree
Agreement	◯	◯	◯	◯	◯

***8. During problem solving,**
I got information from one page confused with information on another page.

	Disagree		Neutral		Agree
Agreement	◯	◯	◯	◯	◯

***9. During problem solving,**
I got lost about where I was in the task (e.g. what page).

	Disagree		Neutral		Agree
Agreement	◯	◯	◯	◯	◯

***10. During problem solving,**
I was trying to get it all done quickly or correctly.

	Quickly		Neutral		Correctly
Rating	◯	◯	◯	◯	◯

Done

APPENDIX G

COGNITIVE LOAD SURVEY

Cognitive Load Survey: part 1

Exit this survey

*1. Name:

*2. How easy or difficult was the task overall? Evaluate the task as a whole.

	Very, very easy (1)	2	3	4	5	6	7	8	Very, very difficult (9)
Rating									

*3. How much mental effort (e.g., thinking, deciding, remembering, looking, searching) did the task take? Evaluate the task as a whole.

	Very, very low mental effort (1)	2	3	4	5	6	7	8	Very, very high mental effort (9)
Rating									

*4. How easy or difficult was the task itself? Evaluate data interpretation task only, excluding display design issue.

	Very, very easy (1)	2	3	4	5	6	7	8	Very, very difficult (9)
Rating									

*5. How much mental effort (e.g., thinking, deciding, remembering, looking, searching) did the task take? Evaluate data interpretation task only, excluding display design issue.

	Very, very low mental effort (1)	2	3	4	5	6	7	8	Very, very high mental effort (9)
Rating									

*6. How easy or difficult was to use the given the display? Evaluate the display design.

	Very, very easy (1)	2	3	4	5	6	7	8	Very, very difficult (9)
Rating									

*7. How much mental effort (e.g., thinking, deciding, remembering, looking, searching) did the display design take? Evaluate the display design.

	Very, very low mental effort (1)	2	3	4	5	6	7	8	Very, very high mental effort (9)
Rating									

Done

Cognitive Load Survey: part 2

Exit this survey

***1. Name:**

***2. How easy or difficult was to understand the theories of the studies?**

	Very, very easy (1)	2	3	4	5	6	7	8	Very, very difficult (9)
Rating									

***3. How much mental effort (e.g., thinking, deciding, remembering, looking, searching) did it take to understand the theories of the studies?**

	Very, very low mental effort (1)	2	3	4	5	6	7	8	Very, very high mental effort (9)
Rating									

***4. How easy or difficult was to understand the methods of the studies?**

	Very, very easy (1)	2	3	4	5	6	7	8	Very, very difficult (9)
Rating									

***5. How much mental effort (e.g., thinking, deciding, remembering, looking, searching) did it take to understand the methods of the studies?**

	Very, very low mental effort (1)	2	3	4	5	6	7	8	Very, very high mental effort (9)
Rating									

***6. How easy or difficult was to understand the results of the studies?**

	Very, very easy (1)	2	3	4	5	6	7	8	Very, very difficult (9)
Rating									

***7. How much mental effort (e.g., thinking, deciding, remembering, looking, searching) did it take to understand the results of the studies?**

	Very, very low mental effort (1)	2	3	4	5	6	7	8	Very, very high mental effort (9)
Rating									

Done

Cognitive Load Survey: part 3

Exit this survey

*** 1. Name:**

*** 2. How easy or difficult was to integrate theories and methods?**

	Very, very easy (1)	2	3	4	5	6	7	8	Very, very difficult (9)
Rating	○	○	○	○	○	○	○	○	○

*** 3. How much mental effort (e.g., thinking, deciding, remembering, looking, searching) did it take to integrate theories and methods?**

	Very, very low mental effort (1)	2	3	4	5	6	7	8	Very, very high mental effort (9)
Rating	○	○	○	○	○	○	○	○	○

*** 4. How easy or difficult was to integrate methods and results?**

	Very, very easy (1)	2	3	4	5	6	7	8	Very, very difficult (9)
Rating	○	○	○	○	○	○	○	○	○

*** 5. How much mental effort (e.g., thinking, deciding, remembering, looking, searching) did it take to integrate methods and results?**

	Very, very low mental effort (1)	2	3	4	5	6	7	8	Very, very high mental effort (9)
Rating	○	○	○	○	○	○	○	○	○

*** 6. How easy or difficult was to integrate results presented in various formats (i.e., tables and graphs) across the two studies?**

	Very, very easy (1)	2	3	4	5	6	7	8	Very, very difficult (9)
Rating	○	○	○	○	○	○	○	○	○

*** 7. How much mental effort (e.g., thinking, deciding, remembering, looking, searching) did it take to integrate results presented in various formats (i.e., tables and graphs) across the two studies?**

	Very, very low mental effort (1)	2	3	4	5	6	7	8	Very, very high mental effort (9)
Rating	○	○	○	○	○	○	○	○	○

Done

APPENDIX H

DEMOGRAPHIC SURVEY

Demographics

Exit this survey

***1. Name:**

***2. Age:**

***3. Year in the program**

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10+

Other (please specify)

***4. Current program:**

✱5. How many years of experience do you have with examining behavioral data (e.g., response time, accuracy, or number of errors)?
Mark the number of years that you have analyzed behavioral data regularly.

- ◯ 1
- ◯ 2
- ◯ 3
- ◯ 4
- ◯ 5
- ◯ 6
- ◯ 7
- ◯ 8
- ◯ 9
- ◯ 10+

✱6. How many years of experience do you have with using ANOVA (Analysis of Variance)?
Mark the number of years that you have used ANOVA regularly.

- ◯ 1
- ◯ 2
- ◯ 3
- ◯ 4
- ◯ 5
- ◯ 6
- ◯ 7
- ◯ 8
- ◯ 9
- ◯ 10+

Done

APPENDIX I

DESCRIPTIVE STATISTICS FOR TIME AND ACCURACY

Table 7. Descriptive statistics for time and accuracy

	Display format							
	Stacked				Distributed			
Expertise level	*N*	*M*	*SD*	*SE*	*N*	*M*	*SD*	*SE*
	Practice time (minutes)							
Novice	30	5.1	1.0	0.2	34	6.1	2.0	0.3
Expert	19	6.1	1.4	0.3	19	5.6	1.1	0.3
	Window time (minutes)							
Novice	30	13.6	4.0	0.7	34	19.2	8.2	1.4
Expert	19	20.1	9.3	2.1	19	15.7	4.5	1.0
	Answering time (minutes)							
Novice	30	12.1	5.3	1.0	34	11.5	4.8	0.8
Expert	19	17.5	8.6	2.0	19	17.1	6.4	1.5
	Total time (minutes) = Window time + Answering time							
Novice	30	25.7	8.1	1.5	34	30.7	10.2	1.7
Expert	19	37.6	14.6	3.4	19	32.8	8.8	2.0
	Total task accuracy (percentages)							
Novice	30	40.0	26.7	5.2	34	27.2	22.5	4.9
Expert	19	44.7	35.9	6.4	19	56.6	28.7	6.4

APPENDIX J

DESCRIPTIVE STATISTICS FOR FIRST PASS FIXATION MEASURES

J.1 COUNT OF FIRST PASS FIXATIONS BY PAGE

Table 8. Descriptive statistics for first pass fixation measures: Count

	Display format							
	Stacked				Distributed			
Expertise level	*N*	*M*	*SD*	*SE*	*N*	*M*	*SD*	*SE*
	Total fixation count							
Novice	29	1410.7	479.2	89.0	32	441.1	414.4	73.3
Expert	19	1560.2	487.8	111.9	19	553.7	215.6	49.5
	Fixation count: Questions							
Novice	29	62.4	45.0	8.4	32	25.2	36.5	6.4
Expert	19	62.1	23.9	5.5	19	21.7	24.5	5.6
	Fixation count: Study 1 Introduction							
Novice	29	156.7	72.8	13.5	32	44.2	69.9	12.4
Expert	19	157.4	55.6	12.7	19	70.4	74.2	17.0
	Fixation count: Study 1 Hypothesis							
Novice	29	96.6	47.8	8.9	32	26.1	41.0	7.2
Expert	19	130.6	83.3	19.1	19	53.7	54.6	12.5
	Fixation count: Study 1 Methods (1)							
Novice	29	127.1	62.2	11.6	32	52.1	75.1	13.3
Expert	19	171.3	70.2	16.1	19	46.7	52.5	12.0
	Fixation count: Study 1 Methods (2)							
Novice	29	78.6	31.6	5.9	32	24.1	37.7	6.7
Expert	19	80.8	25.8	5.9	19	22.6	32.2	7.4
	Fixation count: Study 1 Results							
Novice	29	116.8	63.2	11.7	32	32.9	43.6	7.7
Expert	19	143.2	111.9	25.7	19	62.9	57.8	13.3
	Fixation count: Study 2 Introduction							
Novice	29	134.1	102.3	19.0	32	54.8	67.2	11.9
Expert	19	150.6	91.0	20.9	19	75.0	91.4	21.0

				Fixation count: Study 2 Hypothesis				
Novice	29	125.3	78.0	14.5	32	38.0	59.9	10.6
Expert	19	115.6	59.1	13.6	19	37.9	55.5	12.7
				Fixation count: Study 2 Methods (1)				
Novice	29	140.6	68.6	12.7	32	42.8	65.3	11.5
Expert	19	161.6	67.6	15.5	19	33.5	52.2	12.0
				Fixation count: Study 2 Methods (2)				
Novice	29	154.5	58.3	10.8	32	34.8	71.2	12.6
Expert	19	151.2	96.4	22.1	19	56.9	60.3	13.8
				Fixation count: Study 2 Results (1)				
Novice	29	99.2	45.8	8.5	32	26.7	38.3	6.8
Expert	19	107.4	77.3	17.7	19	35.4	37.4	8.6
				Fixation count: Study 2 Results (2)				
Novice	29	60.4	35.9	6.7	32	17.0	15.6	2.8
Expert	19	64.5	32.0	7.3	19	13.7	17.0	3.9
				Fixation count: Study 2 Results (3)				
Novice	29	58.4	35.1	6.5	32	22.4	18.3	3.2
Expert	19	63.9	38.2	8.8	19	23.3	19.8	4.5

J.2 AVERAGE OF FIRST PASS FIXATIONS BY PAGE

Table 9. Descriptive statistics for first pass fixation measures: Average

	Display format							
	Stacked				Distributed			
Expertise level	*N*	*M*	*SD*	*SE*	*N*	*M*	*SD*	*SE*
	Total average fixation							
Novice	29	264.8	47.2	8.8	32	238.4	46.3	8.2
Expert	19	251.1	45.6	10.5	19	243.2	44.9	10.3
	Average fixation duration (ms): Questions							
Novice	24	233.1	37.7	7.7	29	206.7	56.8	10.5
Expert	19	226.7	42.0	9.6	15	213.9	54.3	14.0
	Average fixation duration (ms): Study 1 Introduction							
Novice	29	252.4	49.7	9.2	31	270.6	153.0	27.5
Expert	19	248.8	47.4	10.9	19	270.2	144.6	33.2
	Average fixation duration (ms): Study 1 Hypothesis							
Novice	28	254.2	51.3	9.7	30	266.5	103.5	18.9
Expert	18	240.1	49.0	11.5	17	261.6	95.4	23.1
	Average fixation duration (ms): Study 1 Methods (1)							
Novice	29	252.7	51.7	9.6	32	227.5	78.3	13.8
Expert	19	246.1	43.1	9.9	19	283.1	274.7	63.0
	Average fixation duration (ms): Study 1 Methods (2)							
Novice	29	260.7	54.2	10.1	29	250.6	121.8	22.6
Expert	19	246.6	50.5	11.6	18	216.8	66.5	15.7
	Average fixation duration (ms): Study 1 Results							
Novice	29	288.9	54.9	10.2	31	250.5	66.2	11.9
Expert	19	272.6	62.5	14.3	19	269.0	78.4	18.0
	Average fixation duration (ms): Study 2 Introduction							
Novice	28	281.7	79.7	15.1	30	253.2	147.7	27.0
Expert	19	240.7	56.3	12.9	19	232.4	63.6	14.6

	Average fixation duration (ms): Study 2 Hypothesis							
Novice	28	250.6	48.4	9.1	30	263.1	125.9	23.0
Expert	18	250.1	53.1	12.5	17	238.7	144.7	35.1
	Average fixation duration (ms): Study 2 Methods (1)							
Novice	28	259.7	69.0	13.0	31	203.3	62.2	11.2
Expert	19	245.4	55.1	12.6	18	253.9	82.1	19.4
	Average fixation duration (ms): Study 2 Methods (2)							
Novice	29	261.9	45.4	8.4	32	235.8	110.4	19.5
Expert	19	258.1	77.6	17.8	19	233.4	53.9	12.4
	Average fixation duration (ms): Study 2 Results (1)							
Novice	29	271.4	53.7	10.0	32	205.4	55.8	9.9
Expert	19	259.1	44.4	10.2	19	234.6	51.9	11.9
	Average fixation duration (ms): Study 2 Results (2)							
Novice	29	276.0	47.2	8.8	32	232.4	89.9	15.9
Expert	19	253.6	49.7	11.4	19	210.9	61.2	14.0
	Average fixation duration (ms): Study 2 Results (3)							
Novice	29	289.4	59.5	11.1	32	240.3	61.6	10.9
Expert	19	274.4	55.4	12.7	19	232.9	61.2	14.0

J.3 SUM OF FIRST PASS FIXATIONS BY PAGE

Table 10. Descriptive statistics for first pass fixation measures: Sum

	Display format							
	Stacked				Distributed			
Expertise level	*N*	*M*	*SD*	*SE*	*N*	*M*	*SD*	*SE*
	Total sum of fixation duration (minutes)							
Novice	29	6.2	2.4	0.4	32	1.8	1.8	0.3
Expert	19	6.6	2.7	0.6	19	2.3	1.0	0.2
	Sum of fixation duration (seconds): Questions							
Novice	29	15.1	12.0	2.2	32	6.0	9.0	1.6
Expert	19	14.7	7.1	1.6	19	5.1	6.5	1.5
	Sum of fixation duration (seconds): Study 1 Introduction							
Novice	29	39.9	20.5	3.8	32	11.3	19.0	3.4
Expert	19	39.3	16.0	3.7	19	16.6	17.8	4.1
	Sum of fixation duration (seconds): Study 1 Hypothesis							
Novice	29	24.7	13.6	2.5	32	6.5	10.8	1.9
Expert	19	32.3	24.8	5.7	19	12.0	12.3	2.8
	Sum of fixation duration (seconds): Study 1 Methods (1)							
Novice	29	32.6	16.7	3.1	32	13.1	20.0	3.5
Expert	19	42.1	18.1	4.2	19	11.5	13.9	3.2
	Sum of fixation duration (seconds): Study 1 Methods (2)							
Novice	29	20.8	9.8	1.8	32	5.7	9.2	1.6
Expert	19	20.2	8.4	1.9	19	5.4	8.0	1.8
	Sum of fixation duration (seconds): Study 1 Results							
Novice	29	34.8	22.4	4.2	32	8.8	12.8	2.3
Expert	19	40.0	33.5	7.7	19	18.7	18.8	4.3
	Sum of fixation duration (seconds): Study 2 Introduction							
Novice	29	34.7	25.7	4.8	32	13.9	17.8	3.2

Expert	19	38.5	25.5	5.9	19	18.4	22.7	5.2
Sum of fixation duration (seconds): Study 2 Hypothesis								
Novice	29	32.0	22.2	4.1	32	10.0	16.3	2.9
Expert	19	29.4	16.9	3.9	19	9.4	14.9	3.4
Sum of fixation duration (seconds): Study 2 Methods (1)								
Novice	29	36.2	21.3	4.0	32	10.3	17.4	3.1
Expert	19	40.1	18.9	4.3	19	9.3	16.9	3.9
Sum of fixation duration (seconds): Study 2 Methods (2)								
Novice	29	41.0	17.7	3.3	32	8.1	17.1	3.0
Expert	19	38.1	25.4	5.8	19	14.8	16.9	3.9
Sum of fixation duration (seconds): Study 2 Results (1)								
Novice	29	27.2	14.4	2.7	32	6.7	10.3	1.8
Expert	19	28.9	23.8	5.5	19	8.9	10.2	2.3
Sum of fixation duration (seconds): Study 2 Results (2)								
Novice	29	16.4	10.2	1.9	32	4.4	4.6	0.8
Expert	19	16.7	10.7	2.5	19	3.4	4.9	1.1
Sum of fixation duration (seconds): Study 2 Results (3)								
Novice	29	17.3	11.9	2.2	32	5.7	4.7	0.8
Expert	19	16.7	9.5	2.2	19	5.7	4.9	1.1

APPENDIX K

DESCRIPTIVE STATISTICS FOR RETURN FIXATION MEASURES

K.1 COUNT OF RETURN FIXATIONS BY PAGE

Table 11. Descriptive statistics for return fixation measures: Count

	Display format							
	Stacked				Distributed			
Expertise level	*N*	*M*	*SD*	*SE*	*N*	*M*	*SD*	*SE*
	Total fixation count							
Novice	29	419.9	338.6	62.9	32	1720.1	903.2	159.7
Expert	19	581.6	330.7	75.9	19	1655.2	575.4	132.0
	Fixation count: Questions							
Novice	29	19.2	37.9	7.0	32	121.9	103.0	18.2
Expert	19	46.2	62.8	14.4	19	96.4	93.9	21.5
	Fixation count: Study 1 Introduction							
Novice	29	6.6	21.8	4.0	32	124.3	103.5	18.3
Expert	19	9.2	28.0	6.4	19	99.8	97.2	22.3
	Fixation count: Study 1 Hypothesis							
Novice	29	22.9	52.3	9.7	32	149.8	144.9	25.6
Expert	19	32.8	53.7	12.3	19	97.1	74.5	17.1
	Fixation count: Study 1 Methods (1)							
Novice	29	15.2	37.3	6.9	32	154.6	117.4	20.7
Expert	19	41.1	53.1	12.2	19	151.2	81.3	18.6
	Fixation count: Study 1 Methods (2)							
Novice	29	35.8	54.8	10.2	32	110.9	94.4	16.7
Expert	19	34.4	35.5	8.1	19	94.2	37.2	8.5
	Fixation count: Study 1 Results							
Novice	29	129.4	129.4	24.0	32	230.3	192.7	34.1
Expert	19	123.5	111.1	25.5	19	160.5	99.1	22.7
	Fixation count: Study 2 Introduction							
Novice	29	14.3	57.5	10.7	32	124.4	123.2	21.8
Expert	19	23.8	55.5	12.7	19	101.5	95.9	22.0

Fixation count: Study 2 Hypothesis								
Novice	29	13.5	23.9	4.4	32	139.1	125.9	22.3
Expert	19	15.7	27.1	6.2	19	128.4	78.0	17.9
Fixation count: Study 2 Methods (1)								
Novice	29	21.1	40.1	7.4	32	139.0	100.4	17.8
Expert	19	29.3	61.8	14.2	19	144.9	95.0	21.8
Fixation count: Study 2 Methods (2)								
Novice	29	46.4	84.4	15.7	32	176.7	119.2	21.1
Expert	19	70.7	86.1	19.8	19	183.4	123.0	28.2
Fixation count: Study 2 Results (1)								
Novice	29	47.1	61.2	11.4	32	121.6	98.7	17.5
Expert	19	46.5	52.8	12.1	19	162.1	72.3	16.6
Fixation count: Study 2 Results (2)								
Novice	29	33.8	41.5	7.7	32	77.8	51.4	9.1
Expert	19	88.3	76.9	17.6	19	173.2	110.5	25.3
Fixation count: Study 2 Results (3)								
Novice	29	14.6	24.8	4.6	32	49.8	60.4	10.7
Expert	19	20.2	31.7	7.3	19	62.5	39.7	9.1

K.2 AVERAGE OF RETURN FIXATIONS BY PAGE

Table 12. Descriptive statistics for return fixation measures: Average

	Display format							
	Stacked				Distributed			
Expertise level	*N*	*M*	*SD*	*SE*	*N*	*M*	*SD*	*SE*
	Total average fixation							
Novice	25	262.2	50.0	10.0	32	268.1	52.7	9.3
Expert	19	256.1	48.8	11.2	19	263.1	45.4	10.4
	Average fixation duration (ms): Questions							
Novice	9	244.5	31.5	10.5	27	262.4	103.1	19.8
Expert	13	239.1	59.5	16.5	15	259.8	64.9	16.7
	Average fixation duration (ms): Study 1 Introduction							
Novice	3	252.1	27.3	15.8	30	288.7	79.1	14.4
Expert	3	240.2	96.3	55.6	19	321.2	120.2	27.6
	Average fixation duration (ms): Study 1 Hypothesis							
Novice	10	248.5	31.0	9.8	27	268.3	51.9	10.0
Expert	9	248.6	66.7	22.2	17	278.1	133.0	32.3
	Average fixation duration (ms): Study 1 Methods (1)							
Novice	7	243.6	67.2	25.4	31	250.9	56.1	10.1
Expert	10	228.9	49.4	15.6	19	239.4	44.1	10.1
	Average fixation duration (ms): Study 1 Methods (2)							
Novice	15	243.5	54.3	14.0	28	252.3	62.8	11.9
Expert	15	242.8	58.9	15.2	18	245.3	49.4	11.6
	Average fixation duration (ms): Study 1 Results							
Novice	22	278.4	74.5	15.9	31	287.0	59.1	10.6
Expert	17	281.0	66.3	16.1	19	280.2	74.3	17.0
	Average fixation duration (ms): Study 2 Introduction							
Novice	3	257.2	25.5	14.7	29	311.3	114.3	21.2
Expert	4	183.7	41.5	20.7	19	304.7	84.1	19.3

	Average fixation duration (ms): Study 2 Hypothesis							
Novice	10	268.8	59.4	18.8	30	269.0	49.4	9.0
Expert	6	225.2	55.9	22.8	17	245.8	45.1	10.9
	Average fixation duration (ms): Study 2 Methods (1)							
Novice	12	274.5	68.4	19.8	31	268.5	81.7	14.7
Expert	9	238.1	61.6	20.5	18	237.9	39.4	9.3
	Average fixation duration (ms): Study 2 Methods (2)							
Novice	14	259.2	70.5	18.8	32	235.3	63.8	11.3
Expert	13	235.1	51.6	14.3	19	239.6	48.3	11.1
	Average fixation duration (ms): Study 2 Results (1)							
Novice	14	277.1	52.4	14.0	31	258.7	45.0	8.1
Expert	13	250.3	49.5	13.7	19	252.4	45.8	10.5
	Average fixation duration (ms): Study 2 Results (2)							
Novice	16	258.7	40.1	10.0	32	257.2	49.6	8.8
Expert	16	278.7	73.0	18.2	19	254.7	44.8	10.3
	Average fixation duration (ms): Study 2 Results (3)							
Novice	11	308.9	61.6	18.6	29	284.2	94.5	17.6
Expert	9	279.6	37.3	12.4	18	242.4	38.4	9.0

K.3 SUM OF RETURN FIXATIONS BY PAGE

Table 13. Descriptive statistics for return fixation measures: Sum

	Display format							
	Stacked				Distributed			
Expertise level	*N*	*M*	*SD*	*SE*	*N*	*M*	*SD*	*SE*
	Total sum of fixation duration (minutes)							
Novice	29	1.9	1.6	0.3	32	7.7	4.3	0.8
Expert	19	2.5	1.6	0.4	19	6.9	2.8	0.6
	Sum of fixation duration (seconds): Questions							
Novice	29	4.8	9.8	1.8	32	30.0	26.3	4.7
Expert	19	11.4	15.3	3.5	19	23.8	24.5	5.6
	Sum of fixation duration (seconds): Study 1 Introduction							
Novice	29	1.6	5.5	1.0	32	35.9	29.0	5.1
Expert	19	2.6	7.6	1.8	19	26.9	25.0	5.7
	Sum of fixation duration (seconds): Study 1 Hypothesis							
Novice	29	5.7	13.2	2.5	32	41.6	41.9	7.4
Expert	19	8.9	15.6	3.6	19	23.1	17.4	4.0
	Sum of fixation duration (seconds): Study 1 Methods (1)							
Novice	29	3.9	10.6	2.0	32	40.4	31.5	5.6
Expert	19	9.8	14.1	3.2	19	35.8	20.8	4.8
	Sum of fixation duration (seconds): Study 1 Methods (2)							
Novice	29	9.1	16.0	3.0	32	29.5	27.3	4.8
Expert	19	8.4	8.4	1.9	19	23.6	11.2	2.6
	Sum of fixation duration (seconds): Study 1 Results							
Novice	29	34.8	38.6	7.2	32	68.1	60.5	10.7
Expert	19	34.8	33.8	7.8	19	42.4	27.8	6.4
	Sum of fixation duration (seconds): Study 2 Introduction							
Novice	29	3.8	15.8	2.9	32	33.7	31.8	5.6

Expert	19	4.8	11.4	2.6	19	27.3	26.1	6.0
		Sum of fixation duration (seconds): Study 2 Hypothesis						
Novice	29	3.6	6.2	1.2	32	37.0	33.0	5.8
Expert	19	3.9	7.4	1.7	19	31.4	20.5	4.7
		Sum of fixation duration (seconds): Study 2 Methods (1)						
Novice	29	6.3	12.9	2.4	32	37.3	31.1	5.5
Expert	19	8.2	18.7	4.3	19	35.6	27.4	6.3
		Sum of fixation duration (seconds): Study 2 Methods (2)						
Novice	29	11.6	21.3	4.0	32	45.0	34.0	6.0
Expert	19	16.8	21.2	4.9	19	43.5	29.3	6.7
		Sum of fixation duration (seconds): Study 2 Results (1)						
Novice	29	12.8	17.1	3.2	32	32.1	27.3	4.8
Expert	19	11.8	13.2	3.0	19	41.2	20.5	4.7
		Sum of fixation duration (seconds): Study 2 Results (2)						
Novice	29	9.0	11.4	2.1	32	20.0	13.5	2.4
Expert	19	23.6	21.0	4.8	19	45.1	31.3	7.2
		Sum of fixation duration (seconds): Study 2 Results (3)						
Novice	29	4.4	7.2	1.3	32	13.4	16.5	2.9
Expert	19	5.6	9.1	2.1	19	16.0	12.1	2.8

APPENDIX L

DESCRIPTIVE STATISTICS FOR OFF SCREEN GAZE MEASURES

L.1 COUNT OF OFF SCREEN GAZE

Table 14. Descriptive statistics for off screen gaze measures: Count

	Display format							
	Stacked				Distributed			
Expertise level	*N*	*M*	*SD*	*SE*	*N*	*M*	*SD*	*SE*
	Off screen gaze duration total							
Novice	29	280.8	150.1	27.9	32	469.0	419.6	74.2
Expert	19	475.4	330.3	75.8	19	334.9	184.6	42.3
	Off screen gaze duration longer than 500ms							
Novice	29	30.1	26.1	4.9	32	68.6	78.0	13.8
Expert	19	78.7	85.9	19.7	19	43.5	47.1	10.8
	Off screen gaze duration longer than 2000ms							
Novice	29	14.2	15.5	2.9	32	33.7	32.9	5.8
Expert	19	41.2	43.8	10.1	19	21.9	24.1	5.5

L.2 AVERAGE OF OFF SCREEN GAZE DURATION

Table 15. Descriptive statistics for off screen gaze measures: Average

	Display format							
	Stacked				Distributed			
Expertise level	*N*	*M*	*SD*	*SE*	*N*	*M*	*SD*	*SE*
	Off screen gaze duration total (ms)							
Novice	29	553.7	516.0	95.8	32	763.5	720.2	127.3
Expert	19	957.8	791.8	181.6	19	723.2	783.7	179.8
	Off screen gaze duration longer than 500ms (seconds)							
Novice	29	3.7	2.6	0.5	32	4.3	3.0	0.5
Expert	19	5.5	5.7	1.3	19	4.0	3.1	0.7

	Off screen gaze duration longer than 2000ms (seconds)							
Novice	26	7.0	4.2	0.8	26	8.3	6.0	1.2
Expert	19	7.4	6.2	1.4	19	6.2	4.0	0.9

L.3 SUM OF OFF SCREEN GAZE DURATION

Table 16. Descriptive statistics for off screen gaze measures: Sum

	Display format							
	Stacked				Distributed			
Expertise level	*N*	*M*	*SD*	*SE*	*N*	*M*	*SD*	*SE*
	Off screen gaze duration total (minutes)							
Novice	29	2.3	2.0	0.4	32	5.5	5.5	1.0
Expert	19	6.5	6.2	1.4	19	3.2	2.5	0.6
	Off screen gaze duration longer than 500ms (minutes)							
Novice	29	1.9	2.0	0.4	32	4.8	4.9	0.9
Expert	19	5.9	6.2	1.4	19	2.7	2.5	0.6
	Off screen gaze duration longer than 2000ms (minutes)							
Novice	29	1.7	1.9	0.4	32	4.2	4.3	0.8
Expert	19	5.2	5.6	1.3	19	2.3	2.1	0.5

L.4 SUM OF OFF SCREEN GAZE DURATION LONGER THAN 2000MS BY PAGE

Table 17. Descriptive statistics for off screen gaze measures: Sum longer than 2000ms

	Display format							
	Stacked				Distributed			
Expertise level	*N*	*M*	*SD*	*SE*	*N*	*M*	*SD*	*SE*
	Sum of fixation duration (seconds): Questions							
Novice	29	6.3	17.4	3.2	32	32.6	77.9	13.8

Expert	19	26.6	52.9	12.1	19	4.3	7.2	1.7
Sum of fixation duration (seconds): Study 1 Introduction								
Novice	29	5.1	10.7	2.0	32	7.5	24.7	4.4
Expert	19	15.8	25.0	5.7	19	4.0	11.2	2.6
Sum of fixation duration (seconds): Study 1 Hypothesis								
Novice	29	9.7	12.6	2.3	32	18.5	30.4	5.4
Expert	19	28.5	32.2	7.4	19	8.6	17.4	4.0
Sum of fixation duration (seconds): Study 1 Methods (1)								
Novice	29	4.5	11.2	2.1	32	17.5	38.1	6.7
Expert	19	12.4	16.5	3.8	19	11.2	17.0	3.9
Sum of fixation duration (seconds): Study 1 Methods (2)								
Novice	29	4.9	11.4	2.1	32	19.2	38.7	6.8
Expert	19	15.5	20.8	4.8	19	10.6	16.8	3.8
Sum of fixation duration (seconds): Study 1 Results								
Novice	29	20.8	29.7	5.5	32	31.6	40.3	7.1
Expert	19	45.1	64.2	14.7	19	22.2	36.1	8.3
Sum of fixation duration (seconds): Study 2 Introduction								
Novice	29	5.3	10.7	2.0	32	8.2	21.7	3.8
Expert	19	29.8	35.5	8.2	19	3.5	10.7	2.5
Sum of fixation duration (seconds): Study 2 Hypothesis								
Novice	29	12.9	17.0	3.1	32	32.5	74.1	13.1
Expert	19	18.0	34.1	7.8	19	10.5	20.5	4.7
Sum of fixation duration (seconds): Study 2 Methods (1)								
Novice	29	3.8	10.2	1.9	32	8.3	17.9	3.2
Expert	19	19.3	24.9	5.7	19	5.4	12.6	2.9
Sum of fixation duration (seconds): Study 2 Methods (2)								
Novice	29	3.9	10.9	2.0	32	36.9	63.6	11.2
Expert	19	32.0	50.6	11.6	19	18.6	42.4	9.7
Sum of fixation duration (seconds): Study 2 Results (1)								
Novice	29	12.9	19.3	3.6	32	11.3	23.3	4.1
Expert	19	26.4	32.5	7.5	19	8.2	13.1	3.0

	Sum of fixation duration (seconds): Study 2 Results (2)							
Novice	29	8.5	17.6	3.3	32	12.3	25.4	4.5
Expert	19	34.9	50.1	11.5	19	9.3	13.5	3.1
	Sum of fixation duration (seconds): Study 2 Results (3)							
Novice	29	0.9	2.2	0.4	32	16.6	38.0	6.7
Expert	19	9.0	14.1	3.2	19	24.9	26.9	6.3

APPENDIX M

NOTE CONTENT ANALYSES

Table 18. Note content analyses

	Display format							
	Stacked				Distributed			
Expertise level	*N*	*M*	*SD*	*SE*	*N*	*M*	*SD*	*SE*
	Total note amount (1: little, 2: light, 3: medium, 4: heavy)							
Novice	29	1.5	1.3	0.2	32	2.3	1.3	0.2
Expert	19	2.6	1.2	0.3	19	1.9	1.2	0.3
	Note amount: Questions							
Novice	21	0.4	1.1	0.2	27	0.1	0.6	0.1
Expert	17	0.2	0.8	0.2	15	0.1	0.3	0.1
	Note amount: Study 1 Introduction							
Novice	21	0.4	0.6	0.1	27	0.3	0.8	0.2
Expert	17	0.7	0.8	0.2	15	0.3	0.6	0.2
	Note amount: Study 1 Hypothesis							
Novice	21	1.1	1.0	0.2	27	0.9	0.7	0.1
Expert	17	1.4	0.9	0.2	15	1.0	0.9	0.2
	Note amount: Study 1 Methods (1)							
Novice	21	0.2	0.4	0.1	27	0.7	1.1	0.2
Expert	17	0.8	0.8	0.2	15	0.8	0.8	0.2
	Note amount: Study 1 Methods (2)							
Novice	21	0.3	0.8	0.2	27	0.7	1.1	0.2
Expert	17	1.1	1.2	0.3	15	0.5	0.9	0.2
	Note amount: Study 1 Results							
Novice	21	1.0	1.1	0.2	27	0.9	1.1	0.2
Expert	17	1.7	1.2	0.3	15	1.9	1.0	0.2
	Note amount: Study 2 Introduction							
Novice	21	0.2	0.7	0.2	27	0.5	0.9	0.2
Expert	17	0.6	1.1	0.3	15	0.3	0.6	0.2
	Note amount: Study 2 Hypothesis							
Novice	21	1.0	0.9	0.2	27	1.0	0.9	0.2
Expert	17	1.1	0.9	0.2	15	0.8	0.9	0.2

				Note amount: Study 2 Methods (1)				
Novice	21	0.2	0.6	0.1	27	0.6	0.8	0.2
Expert	17	0.9	0.9	0.2	15	0.5	0.9	0.2
				Note amount: Study 2 Methods (2)				
Novice	21	0.2	0.6	0.1	27	1.0	1.4	0.3
Expert	17	1.7	1.4	0.3	15	0.8	1.1	0.3
				Note amount: Study 2 Results (1)				
Novice	21	0.6	1.1	0.2	27	0.7	1.1	0.2
Expert	17	1.1	1.1	0.3	15	1.0	0.9	0.2
				Note amount: Study 2 Results (2)				
Novice	21	0.6	0.8	0.2	27	0.7	1.1	0.2
Expert	17	0.8	1.0	0.3	15	0.7	0.9	0.2
				Note amount: Study 2 Results (3)				
Novice	21	0.2	0.5	0.1	27	0.6	1.0	0.2
Expert	17	0.6	0.9	0.2	15	1.0	1.0	0.3
				Number of integrative notes				
Novice	21	0.1	0.3	0.1	27	0.2	0.6	0.1
Expert	17	0.6	0.8	0.2	15	0.4	0.6	0.2
				Number of inferences written				
Novice	21	1.4	1.6	0.3	27	2.3	2.1	0.4
Expert	17	2.4	2.4	0.6	15	2.3	1.5	0.4

APPENDIX N

DESCRIPTIVE STATISTICS FOR SURVEYS

N.1 STRATEGY SURVEY

Table 19. Descriptive statistics for surveys: Strategy survey

	Display format							
	Stacked				Distributed			
Expertise level	*N*	*M*	*SD*	*SE*	*N*	*M*	*SD*	*SE*
	Had trouble remembering info & had to look back							
Novice	30	3.7	0.9	0.2	34	3.9	0.9	0.2
Expert	19	3.8	0.9	0.2	19	3.2	1.3	0.3
	Relied on information on the screen instead of remembering							
Novice	30	3.1	1.2	0.2	34	3.8	0.9	0.2
Expert	19	3.4	1.2	0.3	19	3.2	1.3	0.3
	Had trouble finding the information I need							
Novice	30	1.9	1.1	0.2	34	1.8	1.2	0.2
Expert	19	1.9	1.2	0.3	19	1.3	0.6	0.1
	All the information from different page blurred							
Novice	30	2.0	1.1	0.2	34	1.7	0.9	0.2
Expert	19	1.9	1.1	0.2	19	1.4	0.7	0.2
	Had trouble remembering hypotheses							
Novice	30	2.5	1.4	0.3	34	2.5	1.3	0.2
Expert	19	2.5	1.5	0.3	19	2.0	1.0	0.2
	Purposely tried to keep numbers in my head							
Novice	30	1.7	0.9	0.2	34	1.6	1.0	0.2
Expert	19	1.5	0.9	0.2	19	1.8	1.1	0.3
	Information from one page confused with another							
Novice	30	1.8	0.9	0.2	34	2.1	1.1	0.2
Expert	19	2.1	1.1	0.2	19	1.5	0.7	0.2
	Got lost where I was in the task							
Novice	30	1.9	1.0	0.2	34	1.8	1.1	0.2
Expert	19	1.4	0.9	0.2	19	1.4	0.9	0.2

	Trying to get it done: 1 = quickly, 3 = neutral, 5 = correctly							
Novice	30	3.9	1.1	0.2	34	4.3	0.8	0.1
Expert	19	4.4	0.6	0.1	19	4.5	0.7	0.2

N.2 COGNITIVE LOAD SURVEY

Table 20. Descriptive statistics for surveys: Cognitive load survey

	Display format							
	Stacked				Distributed			
Expertise level	*N*	*M*	*SD*	*SE*	*N*	*M*	*SD*	*SE*
	Task overall: How easy or difficult							
Novice	30	4.0	1.2	0.2	34	5.0	1.7	0.3
Expert	19	4.7	1.6	0.4	19	4.4	1.6	0.4
	Task overall: How much mental effort							
Novice	30	5.2	1.5	0.3	34	5.9	1.4	0.2
Expert	19	5.6	1.6	0.4	19	5.1	1.7	0.4
	Task itself: How easy or difficult							
Novice	30	4.4	1.4	0.2	34	4.8	1.5	0.3
Expert	19	5.1	1.4	0.3	19	3.8	1.5	0.3
	Task itself: How much mental effort							
Novice	30	4.7	1.3	0.2	34	5.8	1.5	0.3
Expert	19	5.3	1.4	0.3	19	4.1	1.4	0.3
	Window design: How easy or difficult							
Novice	30	3.5	2.1	0.4	34	2.3	1.6	0.3
Expert	19	4.0	1.5	0.3	19	2.1	0.9	0.2
	Window design: How much mental effort							
Novice	30	3.9	1.9	0.4	34	2.7	1.7	0.3
Expert	19	3.6	1.6	0.4	19	1.9	0.8	0.2
	Understanding theory: How easy or difficult							
Novice	30	2.9	1.0	0.2	34	3.6	1.8	0.3
Expert	19	2.4	1.0	0.2	19	3.1	1.4	0.3
	Understanding theory: How much mental effort							

Novice	30	3.6	1.4	0.2	34	4.3	1.9	0.3
Expert	19	2.8	1.2	0.3	19	3.4	1.8	0.4
			Understanding methods: How easy or difficult					
Novice	30	3.3	1.6	0.3	34	4.1	1.7	0.3
Expert	19	3.3	1.4	0.3	19	2.8	1.5	0.4
			Understanding methods: How much mental effort					
Novice	30	3.8	1.4	0.3	34	4.6	1.8	0.3
Expert	19	3.7	1.5	0.3	19	3.2	1.8	0.4
			Understanding results: How easy or difficult					
Novice	30	4.8	1.4	0.3	34	5.1	2.0	0.3
Expert	19	4.5	1.5	0.3	19	3.5	1.8	0.4
			Understanding results: How much mental effort					
Novice	30	5.0	1.4	0.3	34	5.6	2.0	0.4
Expert	19	5.2	1.6	0.4	19	3.8	1.6	0.4
			Integrating theory and methods: How easy or difficult					
Novice	30	3.7	1.3	0.2	34	4.2	1.9	0.3
Expert	19	3.4	1.3	0.3	19	3.5	1.6	0.4
			Integrating theory and methods: How much mental effort					
Novice	30	3.9	1.4	0.3	34	4.6	1.9	0.3
Expert	19	3.8	1.5	0.3	19	3.9	1.6	0.4
			Integrating methods and results: How easy or difficult					
Novice	30	4.5	1.5	0.3	34	4.6	2.0	0.3
Expert	19	3.8	1.5	0.3	19	3.5	1.5	0.3
			Integrating methods and results: How much mental effort					
Novice	30	5.1	1.5	0.3	34	5.0	2.1	0.4
Expert	19	4.1	1.4	0.3	19	3.8	1.7	0.4
			Integrating results and results: How easy or difficult					
Novice	30	4.2	1.9	0.3	34	4.7	1.8	0.3
Expert	19	4.2	1.7	0.4	19	3.3	1.8	0.4
			Integrating results and results: How much mental effort					
Novice	30	4.5	1.5	0.3	34	5.0	1.9	0.3

Expert	19	4.5	1.9	0.4	19	3.6	1.5	0.3
			Total cognitive load: Perceived difficulty					
Novice	30	3.9	0.9	0.2	34	4.3	1.3	0.2
Expert	19	3.9	0.9	0.2	19	3.3	1.0	0.2
			Total cognitive load: Mental effort					
Novice	30	4.4	1.0	0.2	34	4.8	1.3	0.2
Expert	19	4.3	1.0	0.2	19	3.6	1.1	0.3
			Total cognitive load					
Novice	30	4.2	0.9	0.2	34	4.6	1.3	0.2
Expert	19	4.1	0.9	0.2	19	3.5	1.0	0.2

APPENDIX O

DESCRIPTIVE STATISTICS FOR PAGE VISITS

Table 21. Descriptive Statistics for page visits

	Display format							
	Stacked				Distributed			
Expertise level	*N*	*M*	*SD*	*SE*	*N*	*M*	*SD*	*SE*
	Total number of visits							
Novice	29	20.6	5.8	1.1	32	156.8	84.3	14.9
Expert	19	24.9	6.0	1.4	19	147.4	39.7	9.1
	Number of visits: Questions							
Novice	29	1.3	1.0	0.2	32	16.8	14.9	2.6
Expert	19	2.3	1.2	0.3	19	11.8	10.1	2.3
	Number of visits: Study 1 Introduction							
Novice	29	1.1	0.3	0.1	32	8.5	6.8	1.2
Expert	19	1.2	0.4	0.1	19	6.5	2.8	0.7
	Number of visits: Study 1 Hypothesis							
Novice	29	1.4	0.8	0.1	32	13.4	12.0	2.1
Expert	19	1.6	1.1	0.2	19	9.9	6.5	1.5
	Number of visits: Study 1 Methods (1)							
Novice	29	1.3	0.7	0.1	32	13.5	10.2	1.8
Expert	19	1.8	1.0	0.2	19	11.1	4.8	1.1
	Number of visits: Study 1 Methods (2)							
Novice	29	1.9	1.2	0.2	32	11.7	8.6	1.5
Expert	19	2.2	1.0	0.2	19	9.7	4.9	1.1
	Number of visits: Study 1 Results							
Novice	29	2.7	1.5	0.3	32	18.3	14.9	2.6
Expert	19	2.4	0.8	0.2	19	11.2	7.3	1.7
	Number of visits: Study 2 Introduction							
Novice	29	1.1	0.4	0.1	32	6.9	5.6	1.0
Expert	19	1.2	0.4	0.1	19	6.3	2.5	0.6
	Number of visits: Study 2 Hypothesis							
Novice	29	1.4	0.7	0.1	32	10.4	9.4	1.7
Expert	19	1.4	0.8	0.2	19	9.0	4.1	0.9

			Number of visits: Study 2 Methods (1)					
Novice	29	1.4	0.7	0.1	32	8.1	6.3	1.1
Expert	19	1.7	0.9	0.2	19	9.2	5.4	1.3
			Number of visits: Study 2 Methods (2)					
Novice	29	1.7	0.9	0.2	32	12.4	7.9	1.4
Expert	19	2.1	1.1	0.3	19	13.9	9.2	2.1
			Number of visits: Study 2 Results (1)					
Novice	29	2.0	1.2	0.2	32	16.1	15.3	2.7
Expert	19	2.2	1.3	0.3	19	18.3	9.2	2.1
			Number of visits: Study 2 Results (2)					
Novice	29	1.8	0.8	0.2	32	13.8	9.4	1.7
Expert	19	3.1	1.7	0.4	19	20.7	13.5	3.1
			Number of visits: Study 2 Results (3)					
Novice	29	1.4	0.6	0.1	32	7.1	6.6	1.2
Expert	19	1.7	1.0	0.2	19	9.7	5.5	1.3

BIBLIOGRAPHY

Breslow, L. A., Trafton, J. G., & Ratwani, R. M. (2009). A perceptual process approach to selecting color scales for complex visualizations. *Journal of Experimental Psychology: Applied, 15*(1), 25.

Carpenter, P. A., & Just, M. A. (1983). What your eyes do while your mind is reading. In K. Rayner (Ed.), *Eye Movements in Reading: Perceptual and Language Processes* (pp. 275-307). New York: Academic Press.

Carpenter, P. A., & Shah, P. (1998). A Model of the Perceptual and Conceptual Processes in Graph Comprehension. *Journal of Experimental Psychology: Applied, 4*(2), 75-100.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 4*(1), 55-81.

Chi, M. T. H. (2006). Laboratory methods for assessing experts' and novices' knowledge. In K. A. Ericsson, N. Charness, P. J. Feltovich & R. R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 167-184). New York, NY, US: Cambridge University Press.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*(2), 121-152.

Chinn, C. A., & Malhotra, B. A. (2002). Epistemologically authentic inquiry in schools: A theoretical framework for evaluating inquiry tasks. *Science Education, 86*(2), 175-218.

Cho, K. (2004). *When experts give worse advice than novices: The type and impact of feedback given by students and an instructor on student writing*. Unpublished dissertation. University of Pittsburgh.

Czerwinski, M., Robertson, G., Meyers, B., Smith, G., Robbins, D., & Tan, D. (2006). *Large display research overview.* Paper presented at the CHI 2006, Montreal, Quebec, Canada.

Freedman, E., & Shah, P. (2002). Toward a model of knowledge-based graph comprehension. *Diagrammatic Representation and Inference*, 59-141.

Gattis, M., & Holyoak, K. J. (1996). Mapping conceptual to spatial relations in visual reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(1), 231.

Gopie, N., & MacLeod, C. M. (2009). Destination Memory: Stop Me if I've Told You This Before. *Psychological Science, 20*(12), 1492-1499. doi: 10.1111/j.1467-9280.2009.02472.x

Gray, W. D., & Fu, W. T. (2004). Soft constraints in interactive behavior: the case of ignoring perfect knowledge in-the-world for imperfect knowledge in-the-head. *Cognitive Science: A Multidisciplinary Journal, 28*(3), 359 - 382.

Gray, W. D., Sims, C. R., Fu, W. T., & Schoelles, M. J. (2006). The soft constraints hypothesis: A rational analysis approach to resource allocation for interactive behavior. *Psychological Review, 113*(3), 461.

Grudin, J. (2001). *Partitioning digital worlds: Focal and peripheral awareness in multiple monitor use.* Paper presented at the CHI 2001: the ACM Conference on Human Factors in Computing Systems.

Hayes, J. R. (1985). Three problems in teaching general skills. *Thinking and learning skills, 2*, 391-406.

Hegarty, M. (2011). The Cognitive Science of Visual‐ Spatial Displays: Implications for Design. *Topics in Cognitive Science, 3*(3), 446-474.

Hegarty, M., Canham, M. S., & Fabrikant, S. I. (2010). Thinking about the weather: How display salience and knowledge affect performance in a graphic inference task. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(1), 37.

Hinsley, D. A., Hayes, J. R., & Simon, H. A. (1977). From words to equations: Meaning and representation in algebra word problems. In M. Just & P. A. Carpenter (Eds.), *Cognitive processes in comprehension* (pp. 89-106). Hillsdale, NJ: Erlbaum.

Hollan, J., Hutchins, E., & Kirsh, D. (2000). Distributed cognition: toward a new foundation for human-computer interaction research. *ACM Transactions on Computer-Human Interaction (TOCHI), 7*(2), 174-196.

Jang, J., & Schunn, C. D. (2008). User satisfaction evaluation and recommendations for Phoenix Science Interface (PSI) in the 2008 Phoenix Mars Mission (Technical Report): University of Pittsburgh, San Jose State University Research Foundation, NASA Ames.

Jang, J., & Schunn, C. D. (2011). Student's adaptive choice of instruction format. In L. Carlson, C. Hölscher & T. Shipley (Eds.), *Proceedings of the 33rd Anuual Conference of the Cognitive Science Society* (pp. 2070-2075): Austin, TX: Cognitive Science Society.

Jang, J., & Schunn, C. D. (2012). Performance Benefits of Spatially Distributed Versus Stacked Information on Integration Tasks. *Applied Cognitive Psychology, 26*(2), 207-214. doi: 10.1002/acp.1811

Jang, J., Schunn, C. D., & Nokes, T. J. (2011). Spatially distributed instructions improve learning outcomes and efficiency. *Journal of Educational Psychology, 103*(1), 60-72. doi: 10.1037/a0021994

Jang, J., Trickett, S. B., Schunn, C. D., & Trafton, J. G. (2012). Unpacking the temporal advantage of distributing complex visual displays. *International Journal of Human-Computer Studies, 70*(11), 812-827. doi: 10.1016/j.ijhcs.2012.07.003

Jones, D. (2010). A WEIRD View of Human Nature Skews Psychologists' Studies. *Science, 328*(5986), 1627. doi: 10.1126/science.328.5986.1627

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist, 38*(1), 23 - 31. doi: 10.1207/S15326985EP3801_4

Kalyuga, S., Chandler, P., & Sweller, J. (1998). Levels of expertise and instructional design (theory on learning). *Human Factors, 40*(1), 1-17. doi: 10.1518/001872098779480587

Kirsh, D. (1995). The intelligent use of space. *Artificial Intelligence, 73*(1), 31-68.

Kirsh, D., & Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cognitive Science, 18*(4), 513-549.

Koriat, A., Ben-Zur, H., & Druch, A. (1991). The contextualization of input and output events in memory. *Psychological Research, 53*(3), 260-270.

Kroft, P. D., & Wickens, C. D. (2002). Displaying multi-domain graphical database information. *Information Design Journal, 11*(1), 44.

Kulatanga-Moruzi, C., Brooks, L. R., & Norman, G. R. (2004). The diagnostic disadvantage of having all the facts: Using comprehensive feature lists to bias medical diagnosis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 563-572.

Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (somtimes) worth ten thousand words. *Cognitive Science, 11*(1), 65-99.

Marsh, R. L., & Hicks, J. L. (2002). Comparisons of target output monitoring and source input monitoring. *Applied Cognitive Psychology, 16*(7), 845-862.

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are Good Texts Always Better? Interactions of Text Coherence, Background Knowledge, and Levels of Understanding in Learning From Text. *Cognition and Instruction, 14*(1), 1 - 43. doi: 10.1207/s1532690xci1401_1

Morgan, P. L., Patrick, J., Waldron, S. M., King, S. L., & Patrick, T. (2009). Improving memory after interruption: Exploiting soft constraints and manipulating information access cost. *Journal of Experimental Psychology: Applied, 15*(4), 291.

Moss, J., Kotovsky, K., & Cagan, J. (2006). The role of functionality in the mental representations of engineering students: Some differences in the early stages of expertise. *Cognitive Science: A Multidisciplinary Journal, 30*(1), 65-93.

Mynatt, E. D., Igarashi, T., Edwards, W. K., & LaMarca, A. (1999). *Flatland: New dimensions in office whiteboards.* Paper presented at the CHI '99, Pittsburgh, PA.

Nie, N., Bent, D. H., & Hull, C. H. (1970). *Spss*: McGraw-Hill.

Novick, L. R., & Catley, K. M. (2007). Understanding phylogenies in biology: The influence of a Gestalt perceptual principle. *Journal of Experimental Psychology: Applied, 13*(4), 197.

Paas, F., & van Merriënboer, J. J. (1994). Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach. *Journal of Educational Psychology, 86*(1), 122-133. doi: 10.1037/0022-0663.86.1.122

Patel, V. L., & Groen, G. J. (1991). The general and specific nature of medical expertise: A critical look. *Toward a general theory of expertise: Prospects and limits*, 93-125.

Peebles, D., & Cheng, P. C. H. (2003). Modeling the effect of task and graphical representation on response latency in a graph reading task. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 45*(1), 28-46. doi: 10.1518/hfes.45.1.28.27225

Ratwani, R. M., Trafton, J. G., & Boehm-Davis, D. A. (2008). Thinking graphically: Connecting vision and cognition during graph comprehension. *Journal of Experimental Psychology: Applied, 14*(1), 36.

Robertson, G., Czerwinski, M., Baudisch, P., Meyers, B., Robbins, D., Smith, G., & Tan, D. (2005). Large display user experience. *IEEE CG&A special issue on large displays, 25*(4), 44-51.

Saariluoma, P., & Kalakoski, V. (1997). Skilled imagery and long-term working memory. *The American Journal of Psychology, 110*, 177-201. doi: 10.2307/1423714

Saariluoma, P., & Kalakoski, V. (1998). Apperception and imagery in blindfold chess. *Memory, 6*(1), 67-90. doi: 10.1080/741941600

Sanfey, A., & Hastie, R. (1998). Does evidence presentation format affect judgment? An experimental evaluation of displays of data for judgments. *Psychological Science, 9*(2), 99-103.

Schunn, C. D., & Anderson, J. R. (1999). The generality/specificity of expertise in scientific reasoning. *Cognitive Science: A Multidisciplinary Journal, 23*(3), 337-370.

Schunn, C. D., & Anderson, J. R. (2001). Science education in universities: Explorations of what, when, and how. In K. Crowley, C. D. Schunn & T. Okada (Eds.), *Designing for Science: Implications from Professional, Instructional, and Everyday Science*. Mawah, NJ: Erlbaum.

Schunn, C. D., & Reder, L. M. (2001). Another source of individual differences: Strategy adaptivity to changing rates of success. *Journal of Experimental Psychology: General, 130*(1), 59.

Schunn, C. D., Saner, L., Kirschenbaum, S., Trafton, J. G., & Littleton, E. B. (2007). Complex visual data analysis, uncertainty, and representation. *Thinking with data*, 27-64.

Shah, P., & Carpenter, P. A. (1995). Conceptual limitations in comprehending line graphs. *Journal of Experimental Psychology: General, 124*(1), 43.

Shah, P., & Freedman, E. G. (2011). Bar and Line Graph Comprehension: An Interaction of Top- Down and Bottom- Up Processes. *Topics in Cognitive Science, 3*(3), 560-578.

Spilich, G. J., Vesonder, G. T., Chiesi, H. L., & Voss, J. F. (1979). Text processing of domain-related information for individuals with high and low domain knowledge. *Journal of Verbal Learning and Verbal Behavior, 18*(3), 275-290.

Sweller, J., van Merriënboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*(3), 251-296. doi: 10.1023/A:1022193728205

Tan, D. S., Gergle, D., Scupelli, P. G., & Pausch, R. (2003). *With similar visual angles, larger displays improve spatial performance.* Paper presented at the CHI 2003, Ft. Lauderdale, Florida.

Tan, D. S., Gergle, D., Scupelli, P. G., & Pausch, R. (2004). *Physically large displays improve path integration in 3D virtual navigation tasks.* Paper presented at the CHI 2004: the ACM Conference on Human Factors in Computing Systems.

Thomas, J. J., & Cook, K. A. (2005). Illuminating the path: The research and development agenda for visual analytics. 2005. *IEEE Computer Society*.

Trafton, J. G., Kirschenbaum, S., Tsui, T. L., Miyamoto, R. T., Ballas, J. A., & Raymond, P. D. (2000). Turning pictures into numbers: extracting and generating information from complex visualizations. *International Journal of Human-Computer Studies, 53*(5), 827-850.

Trafton, J. G., Marshall, S., Mintz, F. E., & Trickett, S. B. (2002). Extracting explicit and implicit information from complex visualizations. In M. Hegarty, B. Meyer & N. H. Narayanan (Eds.), *Diagramatic representation and inference*. Berlin Heidelberg: Springer-Verlag.

Trafton, J. G., & Trickett, S. B. (2001). Note-taking for self-explanation and problem solving. *Human-computer interaction, 16*(1), 1-38.

Trafton, J. G., Trickett, S. B., Schunn, C. D., & Kirschenbaum, S. (2007). Complex visualizations, spatial transformations, and uncertainty. Washington, D.C.: Presented at the meeting of the Naval Research Laboratory.

Trickett, S. B., Trafton, J. G., Saner, L., & Schunn, C. D. (2007). "I don't know what's going on there'': The use of spatial transformations to deal with and resolve uncertainty in complex visualizations. In M. Lovett & P. Shah (Eds.), *Thinking with data* (pp. 65-85): Lawrence Erlbaum Associates.

Wellman, B. (1998). Doing it ourselves: The SPSS manual as sociologyís most influential recent book. *Required reading: Sociologyís most influential books*, 71-78.

Willows, D. M. (1978). A picture is not always worth a thousand words: Pictures as distractors in reading. *Journal of Educational Psychology, 70*(2), 255.

Yeh, M., & Wickens, C. D. (2001). Attentional filtering in the design of electronic map displays: A comparison of color coding, intensity coding, and decluttering techniques. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 43*(4), 543-562.

Yeung, A. S., Jin, P., & Sweller, J. (1998). Cognitive Load and Learner Expertise: Split-Attention and Redundancy Effects in Reading with Explanatory Notes. *Contemporary Educational Psychology, 23*(1), 1-21. doi: 10.1006/ceps.1997.0951

Zhang, J., & Norman, D. A. (1994). Representations in distributed cognitive tasks. *Cognitive Science, 18*(1), 87-122.