5-15-2015

Data-ink Ratio and Task Complexity in Graph Comprehension

Kevin McGurgan

Department of Psychology, College of Liberal Arts

Rochester Institute of Technology

Data-ink Ratio and Task Complexity in Graph Comprehension

A Thesis in Experimental Psychology

by

Kevin McGurgan

Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Science

May 15, 2015

We approve the thesis of Kevin McGurgan:

Date

Dr. Andrew Herbert

Professor/Departmental Chair, Dept. of Psychology, RIT

Date

Dr. Elena Fedorovskaya

Professor, College of Imaging Arts & Sciences, RIT

Date

Dr. Tina Sutton

Assistant Professor, Dept. of Psychology, RIT

Acknowledgements

First and foremost, I'd like to thank my thesis advisor, Andrew Herbert, and my thesis committee, Elena Fedorovskaya and Tina Sutton, for their support throughout this project. I'd like to thank the faculty and staff in the Department of Psychology, who have provided support in many ways. Particular thanks go to those professors who allowed me to collect data in their classes. I'd like to thank the College of Liberal Arts for providing. And finally, I'd like to thank the professors who graciously agreed to be interviewed as part of this project.

Abstract

Human processing of graphical information is a topic which has wide-reaching implications for decision-making in a variety of contexts. A deeper understanding of the processes of graphical perception can lead to the development of design guidelines which can enhance performance in graphical perception tasks. This study evaluates the data-ink ratio guideline, which recommends the removal of non-data graph elements, resulting in minimalist graph designs. In an experiment, participants answered graph comprehension questions using bar graphs and boxplots with varying data-ink ratios. Participants answered questions with similar levels of accuracy and mental effort. Some participants drew on graphs, reducing the data-ink ratio of high and medium data-ink stimuli. Additionally, expert interviews were conducted regarding graph use, graph creation, and opinions about the data-ink concept and example graphs. Interviewees had a variety of opinions and preferences with regard to graph design, many of which were dependent upon the specific circumstances of presentation. Most interviewees did not think that high data-ink graph designs were superior. These results suggest that data-ink maximization does not improve performance in graph comprehensions tasks, and that arguments regarding the data-ink ratio deal with the subjective issue of graph aesthetics.

Contents

List of Figures

List of Tables

Data-ink Ratio and Task Complexity in Graph Comprehension

The job of graph designers is to present information in a way which viewers can understand and use meaningfully (Katz, 2012). Achieving this goal requires not only good intentions on the part of designers, but scientifically-derived models which identify the processes that characterize "understanding." Ultimately, the utility of research on graphical comprehension lies in its ability to inform design decisions and aid in the creation of design guidelines.

Graphs are defined as typically static paper or electronic representations of numeric analog data with multiple data points (Wickens & Holland, 2000). They provide a means of communicating quantitative information in an easily-comprehensible format and can make complex information visually salient (Shah, Freedman, & Vekiri, 2005). Graphs allow people to perform tasks that would be more difficult using *other* formats by shifting some of the cognitive demands to the visual system (Lohse, 1997). Graphs have also been said to reduce demands on memory, to group related information for easy search, and to reduce the complexity of tasks by imposing structure on data (Tory & Moller, 2004). However, poorly designed graphs can lead to difficulty in understanding information and ultimately to negative consequences (Freedman & Shah, 2002).

One widely-cited graph design guideline is the *data-ink ratio*, proposed by Edward Tufte, statistician and information design pioneer. Tufte argues that because the purpose of a graph is to help people reason about data, graphs should draw viewers' attention to the data, and not to something else; his fundamental principle for the design of good graphs is to "above all else show the data." (Tufte, 2001, p. 92). The data-ink ratio concept suggests that graph designers should remove elements which do not depict statistical information, resulting in minimalist-style graph designs (Tufte, 2001). Despite widespread acceptance of this design principle, it lacks empirical validation, a persistent problem in the field of information visualization (Zhu, 2007).

Models of Graphical Comprehension

Generally speaking, graph comprehension is a type of human information processing (Wickens & Holland, 2000). Models of human information processing provide a framework for understanding the psychological processes involved in a given task, as well as a means of understanding why performance in a task might change under differing conditions. One such model of information processing is presented in Figure 1 below.

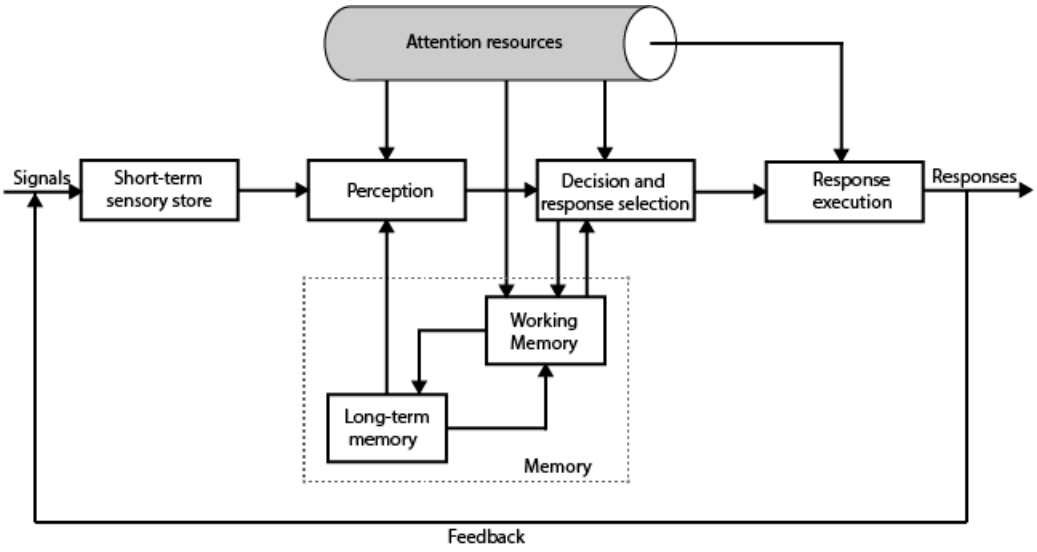


Figure 1. A model of human information processing, adapted from Wickens & Holland (2000).

The model begins with sensory processing, the stage in which environmental information is gathered by sensory organs. This stage of processing is affected by the quality of information from the environment and the short-term memory stores associated with each sensory system. Next comes perception, where raw sensory information is interpreted for use by the brain. This is impacted by both bottom-up processing, or features of the sensory information itself, and top-down processing, in which experience-based information from long-term memory affects interpretation. Next are cognitive process, which are slower and less automatic than perceptual processes. These include reasoning, rehearsal and mental transformation, and, because they also require

working memory, are resource-limited. Material that is sufficiently rehearsed during this stage can be entered into the more permanent long-term memory. Once the previously-described stages of information processing contribute to an overall understanding of a situation or stimuli, response selection occurs. This frequently, though not necessarily, results in the execution of said response. Finally, feedback for an executed response is received from the environment. Feedback is often, though not always, received continuously during the course of a task. It is important to note that the limited availability of attentional resources affects nearly *all* of the stages of processing (Wickens & Holland, 2000).

Based on the human information processing framework, models specific to the processes of graphical comprehension have also been developed. An early model, which focused primarily on perceptual processes, was developed by Cleveland and McGill (1985). They identified ten tasks, described as "elementary graphical-perception tasks" (Cleveland & McGill, 1985, p. 828), and ranked them according to the accuracy with which they can be judged. They include angle, length, position along a common scale, and positions on identical but nonaligned scales, among others. These are distinct from cognitive tasks, they argue, in that our preattentive visual system can detect geometric patterns and assess magnitudes. According to the model, selecting design options higher in ranking should increase the accuracy of perceptions of patterns in graphical data. So, for example, graphs which rely on positioning items along a common scale to display data will be more accurately perceived than graphs which rely on differences in area (Cleveland & McGill, 1985).

Although Cleveland and McGill's (1985) model emphasizes early perceptual processing, it does not speak to other factors important to human information processing (Carswell, 1992). To address this concern, Carswell (1992) conducted a meta-analysis of 39 experiments involving graphical perception tasks in which graphical format was a primary independent variable and dependent variables included at least one measure of

performance, such as response time or accuracy. Task variables ranged from identifying a single data point to integrating most or all of the data points in a given graph. Overall, the elementary tasks model was found to predict performance better than Tufte's data-ink ratio. However, when broken down by task type, it was shown that the predictions of the elementary task model actually contradicted results of studies which involved information synthesis (those that involved comparing most/all of the data points). Carswell (1992) concluded that this model is most successful in predicting performance in local comparison and point-reading tasks, and less successful in global comparison, or synthesis, tasks. This suggests that although the elementary tasks model is a good predictor of performance in certain graphical comprehension tasks, task demands are also an important feature of graph comprehension.

Simkin and Hastie (1987) developed a model of graphical comprehension which also accounted for other stages of information processing. Their research focused on the idea that the utility of a given graph depends on the interaction between the particular design elements of that graph and the tasks being performed by the graph viewer, and identified processes of graph comprehension. For example, *anchoring* is described as isolating a component of the graph to determine an initial value that serves as a standard for an estimate. *Scanning* is described as "sweeping across" the distance between graph elements to make a value estimation; the duration of the scan may inform the estimation. *Projection* is described as "sending out a ray" (Simkin & Hastie, 1987, p. 460) from one point in a graph to another to facilitate comparisons. *Superimposition* is the mental movement of an element so as to overlap with another element and allow for comparison. Finally, *detection operators* detect size differences in graph elements, and result in dichotomous "larger vs. smaller" judgments.

One of the earliest attempts at a comprehensive model of graphical comprehension was made by Pinker (1990). Like Cleveland and McGill (1985), he argued that graph viewers first process the raw sensory data of a graph, which he refers to as a *visual array*.

Using visual encoding processes, a *visual description* of relevant features, such as the shapes used in a graph, is formed. Visual descriptions, he argued, are constrained by features of our vision system, such as processing capacity. Graph viewers recognize the type of graph they are seeing through *instantiated graph schema*, or prior knowledge about specific graph types. He also argues that people create instantiated schemas for specific types of graphs based on a *general graph schema*, which includes knowledge about what graphs are typically used for and how they are typically interpreted. Graph schemas can be enriched through multiple means, including formal instruction, practice at reading graphs, as well as experience with graph creation (Pinker, 1990).

Pinker emphasizes the importance of the *gestalt laws* of perception, which describe the preattentive grouping of visual elements by the perceptual system (Levitin, 2011). Gestalt psychologists asserted that psychological phenomenon need to be understood as organized, structured wholes. These laws act as a constraint on graph comprehension (Shah, 1995), as our visual system automatically groups visual input into distinct psychological units based on visual properties (Kosslyn, 2006). For example, the law of proximity states that objects that are close to each other will tend to be seen as a group. A graph designer may vary the distance between graphical elements to imply relationships between subsets of those elements – elements that are close together tend to be perceived as belonging to the same group or object. The principle of similarity states that objects with similar properties, such as color or shape, will be grouped together. Graph designers frequently use color to associate and/or differentiate graphical elements, and effective use of color can lead to grouping of elements, even when they are spatially separated (Wickens & Holland, 2000). The principle of closure refers to the tendency of the perceptual system to "close" or "fill in" missing parts of objects. For example, a graph designer might include gridlines which fall behind the bars in a graph. Although the bars break the continuity of the gridlines, they are perceived as continuous nonetheless. In general, gestalt psychologists argued that all of the gestalt laws are examples of the *law of pragnanz* ("good figure"), which states

that we perceive the simplest organization which fits the stimulus (Levitin, 2011).

It has been suggested that leveraging the gestalt laws in graph design can lead to faster and more accurate graph comprehension because they exploit the fast-working visual system as opposed to slower, more deliberative cognitive systems (Kirk, 2012). Grouping is automatic and preattentive, and will occur regardless of the designer's intent (Robertson, Czerwinski, Fisher, & Lee, 2009). Shah, Mayer and Hegarty (1999) found that effective use of gestalt grouping in graph design *can* influence viewers' ability to recognize trends in data. Displays designed with effective grouping are said to be highly redundant, meaning that knowledge of the location of one element of the display will facilitate accurate inferences regarding the location of other display elements (Wickens & Holland, 2000). This suggests that effective design requires appropriate use of gestalt grouping principles, provided that it supports the particular task of the viewer.

Lohse (1997) worked towards a better understanding of the role of working memory in graphical comprehension, a bottleneck in information processing with important implications for graph comprehension. He found that participants with high working memory capacity who used a single-color graph were able to make decisions as accurately as participants who used color-enhanced graphs, and concluded that different participants were able to make the same judgments with different levels of efficiency, but that the effect is mediated by task complexity. In other words, individual cognitive limitations exist, regardless of graph design, and different designs cannot result in inherent improvements in complex tasks. However, design features which improve peoples' ability to process information in parallel can result in the distribution of some of the cognitive burden to perceptual systems. That frees working memory resources for other aspects of the comprehension process, requiring less overall effort on the part of the graph viewer. Furthermore, Lohse concludes that quantifying the amount of effort required to extract information from graphs is important in designing information which matches users' goals and is a promising direction for future studies.

Freedman and Shah (2002) developed what they refer to as an "interactive" model of graphical comprehension. This model proposes that graphical comprehension is a sequential constraint-satisfaction process in which characteristics of the viewer, such as prior knowledge, expectations, and graphical literacy skills can impact interpretation. Expertise plays an important role in this model, as experts can automatically relate the visual features of a graph to subject matter knowledge and theoretical interpretations. They also propose that graph information is sequentially encoded in *visual chunks*. The manner in which chunks are perceptually grouped can be affected by visual characteristics – namely gestalt grouping principles, such as the use of varying distances between bars in a bar graph – and effective grouping has been shown to improve viewers' ability to identify trends (Shah et al., 1999). However, deficits in graph skills or domain knowledge can make comprehension more effortful (Freedman & Shah, 2002).

In a review of models of graphical comprehension, Freedman and Shah (2002) identified five factors that interact to influence graphical comprehension: display characteristics of the graph; level of complexity of the data; the specific task of the viewer; the viewer's prior knowledge about the content of the graph; and the viewer's knowledge about graphs in general. They do not, however, identify the relative importance of these five factors with regard to graph comprehension or describe the specific interactions which can occur between them.

The Data-Ink Ratio

Tufte proposes that there are two types of information in a graph – *data-ink* and *non-data-ink*. Data-ink is "the non-erasable core of a graphic" and "the non-redundant ink arranged in response to variation in the numbers represented" (Tufte, 2001, p. 93). According to Tufte, all ink that does not depict statistical information, or "chartjunk," should be removed, and the reason to include additional ink in a graph should almost always be that it depicts *new* information. In other words, most of the ink in a graph "should vary in response to data variation" (Tufte, 2001, p. 136) and data-ink ratios should

be maximized. Tufte notes that chartjunk should only be removed "within reason" (Tufte, 2001, p. 96), though it is unclear what he considers to be reasonable. He also states that data-ink maximization makes sense for roughly two-thirds of graphs, but specific guidelines for determining whether a graph qualifies are absent (Tufte, 2001).

Data-ink ratios range between zero and one and can be calculated by dividing data-ink by the total amount of ink in a graphic (Tufte, 2001). However, it is unclear how a data-ink ratio can be accurately calculated in practice, and Tufte himself seems to make estimations rather than numerical calculations. Figures 2 and 3 provide examples of graphs with varying data-ink ratios. Tufte argues that his redesigned boxplot is particularly suited for exploratory data analysis by researchers because it would take less time to create, though this point may be less relevant and even counter-intuitive given the ubiquity of modern graph-making software today.

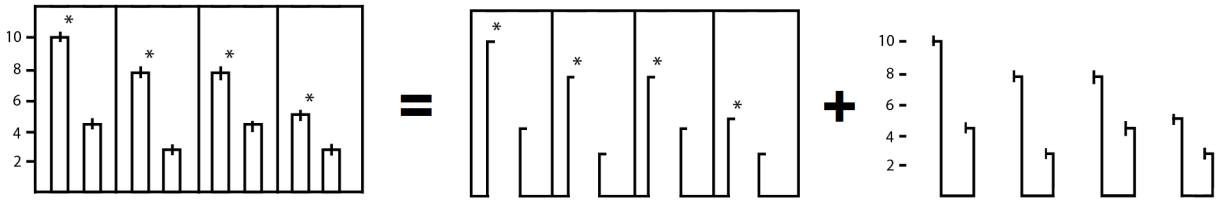


Figure 2. An illustration of erasable non-data-ink adapted from Tufte (2001). On the left is the original design, in the middle is what he recommends erasing, and on the right is the final design, which Tufte argues is an immense improvement over the original.

Tufte goes on to argue that he doesn't believe his unconventional designs would confuse new viewers and that it would be a mistake to underestimate the abilities of audiences of graphical information. He believes that his designs could look strange at first, but would be accepted over time (Tufte, 2001). Tufte seems to make the assumption that graph readers will be able to apply their *instantiated graph schema* to his graph designs. Whether or not this is the case is unclear; for example, it is possible that a graph reader's instantiated boxplot schema may *not* help them to recognize Tufte's redesigned boxplot, as

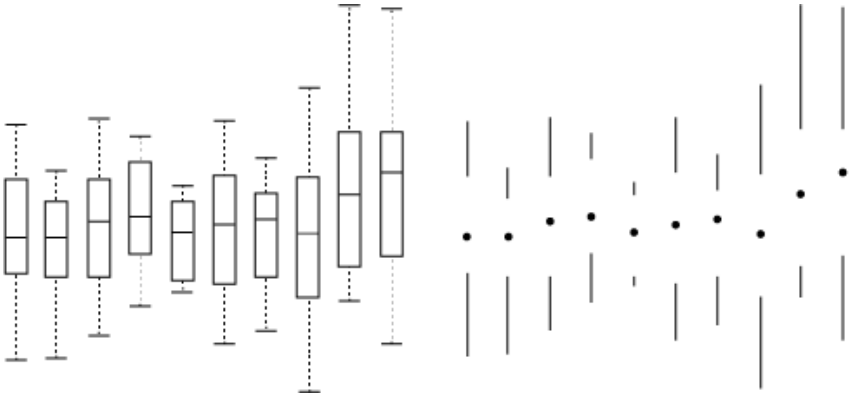


Figure 3. Standard boxplots (left) and Tufte's proposed high data-ink redesign (right)

it lacks the "box" portion of the plot which is usually included. At the same time, graph comprehension theories suggest that it is possible that instruction or extended exposure to high data-ink graph designs *could* cause them become incorporated into graph readers' existing schemas.

With regard to the graph comprehension processes described by Simkin and Hastie (1987), one could argue that high data-ink designs will make comprehension tasks more difficult by removing the elements that viewers use to complete them. For example, by replacing the box in a boxplot with negative space, viewers may have trouble superimposing the negative space of an individual boxplot on top of another to compare the size of those features. Similarly, it could be argued that the inclusion of gridlines facilitates projection, or "sending out a ray," from one element to another – rather than requiring viewers to imagine horizontal lines, they are provided in the design of the graph.

Generally speaking, the data-ink ratio is an influential concept in the field of design (Zhu, 2007; Fry, 2008), and it is believed that higher data-ink ratios will result in faster judgments and increased accuracy in graph reading tasks (Wickens & Holland, 2000). However, others have characterized the data-ink ratio as having its basis in Tufte's design intuitions and lacking experimental validation with behavioral data (Carswell, 1992).

Responses to the Data-ink Ratio

Scholars have published mixed opinions regarding the data-ink ratio concept. Wainer (1984) argues that it is a convenient way to measure the extent to which "chartjunk" is used, and that the closer to zero the ratio gets, the worse the graph. He also argues against the use of additional dimensions and "worthless metaphors," such as showing dollar bills of varying size to indicate changes in purchasing power with time.

On the other hand, Tukey (1990) describes the data-ink ratio as a "dangerous idea" and argues that overreliance on it can be destructive and result in graphs that are both busy and distracting, such as the box plot on the right in Figure 4. Tukey argues that, although the "open" boxplot in Figure 4 has a lower data-ink ratio, the removal of ink draws unnecessary attention to its incompleteness, causing the viewer to wonder why a line is missing. This criticism seems to make more sense in the context of a boxplot with multiple elements, such as those in Figure 2. It could be argued that the removal of the box portion of graph elements results in three distinct perceptual groupings – the upper set of lines, the median dots, and the lower set of lines. Tufte's redesigned boxplot is "destructive" in the sense that visually similar elements of the set of boxplots are grouped with each other, rather than as a part of an individual boxplot element. Kosslyn advocates for additional ink in graphs when it "completes a form, resulting in fewer perceptual units" (Kosslyn, 2006, p. 13.), an argument which seems to draw upon the gestalt grouping principles, and which would recommend *against* high data-ink designs that result in fragmentary graph elements.

Additionally, Tukey describes the process of graph comprehension as requiring attention shifts between the various elements of a graph and argues that individual graph elements should have equal visual impacts. He describes "well-tuned" graph design as a balance between the inclusion of helpful elements and the elimination of features which contribute to busyness. Tukey notes that the purpose of a boxplot's "box" is to emphasize the central clumping of a data distribution, and questions whether the replacement of the

"boxes" in a boxplot with white space, as Tufte recommends, is "strong enough" to facilitate these attention shifts. Tukey's idea of "attention shifts" could be seen as a general way of describing Simkin and Hastie's (1987) comprehension processes. In that case, Tukey would seem to agree that removing elements from a graph would increase the effort required to read that graph, as it would be more difficult to shift attention between two or more graph elements when those elements are small or removed altogether. Tukey concludes that the *underlying* idea behind data-ink ratio might be to avoid busyness and distraction in graph design, and agrees with both of those principles, but points out that those recommendations *alone* would not produce the type of graphs which Tufte advocates (Tukey, 1990).

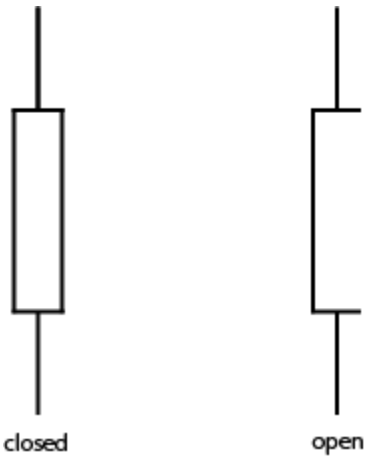


Figure 4. Adapted from Tukey (1990). He argues that although the open symbol on the right has a lower data-ink ratio, it is distracting and directs attention to the right for no purpose.

Others have taken a more nuanced view of the data-ink ratio concept. Kosslyn (1985) found that maximizing data-ink ratio produces the desirable outcome of eliminating potentially distracting decorations, but argues that it is not clear *how* to select nondata-ink and finds Tufte's recommendations to be too extreme overall. Carswell also agrees that "distinguishing data-ink from erasable ink is a subjective endeavor" (Carswell, 1992, p. 540).

Specifically, Kosslyn disagrees with Tufte's recommendation to eliminate horizontal gridlines on bar graphs and to instead use horizontal "lines" created by negative space across the bars (Figure 5) because the perceptual system fills in the "virtual lines," which is just as distracting as standard grid lines (Kosslyn, 1985). This is slightly different from Tukey's criticism of the data-ink ratio on the grounds that it is destructive – Kosslyn argues that, from a perceptual standpoint, negative space lines are the same as standard gridlines, and that both gridline styles can be distracting. Kosslyn does not offer much of an argument as to why or in what circumstances gridlines are distracting, but in later publications argued that gridlines *can* in fact be useful, provided they are not too heavy and do not obscure graph content (Kosslyn, 2006). It could be argued, however, that Tufte's virtual lines *do* in fact require more work on the part of viewers, as the perceptual system completes the virtual lines rather than the lines being completed via ink. From a practical standpoint, the negative space lines in Figure 5 could also cause problems when values fall in the small region of the each bar where the white lines fall. This type of design could feasibly lead to inaccurate graphs.

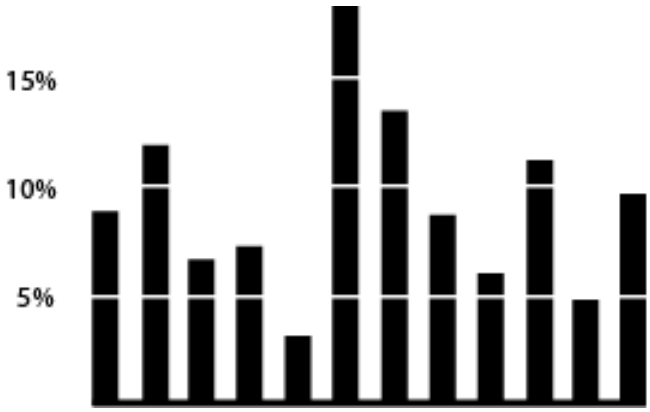


Figure 5. An adapted version of Tufte's proposed bar graph redesign, which Kosslyn argues is no better than a "standard" bar graph.

Stephen Few (2009), data visualization expert, also disagrees with aspects of data-ink ratio. Though he agrees that decorative elements and non-data ink which serve no meaningful purpose should be removed, he believes that some non-data ink, such as axis

lines, should be kept, but with a reduced salience as to not compete with the data itself for viewers' attention. He also argues that some redundancy in graphs is useful and takes issue with the style of many of Tufte's high data-ink graphs on the grounds that reducing the size or visual weight of graph features forces viewers' eyes to "work too hard" and slows them down (Few, 2009, p. 3). Though it is not clear precisely what Few means in psychological terms, the idea that less salient visual objects are more difficult to see is uncontroversial, and it can be assumed that Few would predict that graphs with less salient features would require more effort. This idea has been referred to as the *discriminability* principle – graph features must be large enough that they can easily be distinguished from the background (Kosslyn, 2006). It is certainly plausible that by reducing the salience of graph features, high data-ink designs shift *less* of the cognitive demand of graph comprehension to the visual system, resulting in increase comprehension effort. As previously discussed, Tufte's "virtual lines" could add to this increase in effort. Few concludes that data-ink should not be increased beyond "the minimum that's required," but *should* be increased when it reduces effort for viewers (Few, 2009, p. 8).

Some have gone as far as to suggest that "chartjunk" or "visual difficulties" may actually *benefit* viewers, and that graphs should *not* be designed to maximize the data-ink ratio, but instead to engage viewers to actively process graphical information. In other words, low data-ink ratios may be useful if the extra ink is used to "personalize, aestheticize, or otherwise make the visualization more enticing to end-users so as to drive more intrinsic desires to engage" (Hullman, Adar, & Shah, 2011, p. 2217). In this view, effectiveness of visualizations is seen as a tradeoff between efficiency and desirable visual difficulties which encourage learning (Hullman et al., 2011).

Empirical tests of the Data-ink Ratio

Empirical tests of the efficacy of maximizing data-ink ratio have also yielded mixed results. Using graphs taken from Tufte (1983), Inbar, Tractinsky and Meyer (2007) attempted to evaluate opinions toward high data-ink ratio graphs among 87 undergraduate

students. In one condition, participants were shown either a "standard" bar graph or a minimalist, high data-ink bar graph. The standard bar graph was overwhelmingly preferred. In another condition, participants were asked to perform several undisclosed data extraction tasks using the Tufte-style graph in attempt to familiarize them with high data-ink design. Still, participants overwhelmingly preferred the standard bar graph design. In a third condition, participants were shown two additional graphs, resulting in a four-graph continuum from low to high data-ink. It was thought that these additional graphs may help participants become more familiar with the idea of Tufte's minimalist designs. Preferences were largely split between the standard bar graph and the second-most minimal graph, suggesting that people *can* be receptive to minimalist designs. However, *no* participants preferred the most minimalist graph and it was rated as significantly less clear than the other three. This suggested that there may be a "sweet spot" for data-ink levels which lies between standard bar graphs and Tufte's recommendations (Inbar et al., 2007, p. 188), though it is unclear whether the study's "familiarization" methods were sufficient, and no objective behavioral data were collected. It is also unclear that simply measuring viewers' graph preferences relates to performance in extracting useful interpretations from graphs with different designs.

Gillan and Sorensen (2009) explored the idea that chartjunk may *improve* performance in certain circumstances using graphs with background images which differed from a graph's "indicator features" (i.e., circular background images with rectangular bar graph elements) leading to a "pop out" effect. Participants were shown graphs with matching backgrounds/indicator features, opposing backgrounds/indicator features (rectangular indicators with circular background elements or vice versa), or no background. Finally, participants were asked two types of questions – difference ("What is the difference between the number of loans for undergraduate students between 1950 and 1990?") and comparison ("in which year do undergraduate students have more loans, 1950 or 1990?")(Gillan & Sorensen, 2009, p. 1097). Comparison questions were universally

answered with nearly perfect accuracy, regardless of graph type. For difference questions, it was found that the presence of a background did reduce accuracy, but only when the features of the background were similar to indicator features. They argue that these results suggest that reducing chartjunk doesn't necessarily improve comprehension, and conclude that Tufte's data-ink ratio theory is too simplistic. It should be noted that the study used simple tasks, asking participants to compute the difference between two data points or determine which value is higher or lower. Participants were not required to identify overall trends in the graph or make predictions using the graphs. There were only 18 participants (Gillan & Sorensen, 2009).

Other findings have suggested that accuracy in recalling descriptive elements of visually embellished graphs is no worse than for plain graphs, both immediately after viewing and after a two to three week delay. Bateman et al. (2010) selected 14 graphs from Holmes' book, all of which contained chartjunk and had low data-ink ratios. For example, one graph (titled "Monstrous Costs") showed House and Senate campaign expenditures over time. The graph itself is a bar graph in in the mouth of a monster, with the monster's teeth acting as bars. Corresponding plain versions of the Holmes graphs were designed with the idea of Inbar's (2007) "sweet spot" in mind. Although the plain versions certainly have a higher data-ink ratio than Holmes-style graphs, they do not closely resemble the examples that Tufte (2001) himself created; the high data-ink ratio graphs used in the study are in fact closer to Tufte's *starting* point than his suggested re-designs (Figure 6).

Twenty participants were shown one of each of the 14 graph pairs – either Holmes-style or plain) and asked four questions per graph requiring them to read and describe the contents, such as "What is the graph about?" or "What is the basic trend of the graph?". No differences in the quality of description, based on a scoring which accounted for accuracy, were found. Next, participants were placed into one of two recall conditions – immediate or long-term. No difference was found for recall between Holmes-style and plain graphs in the immediate recall condition, which began five minutes

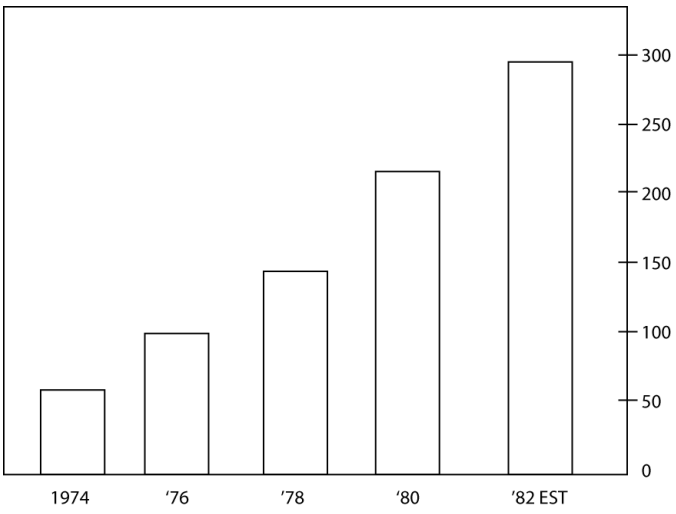


Figure 6. Adapted version of the plain graphs used by Bateman et al. (2010), which did not resemble Tufte's recommended high data-ink ratio bar graph design.

after initial questioning. However, in the long-term recall group, participants remembered significantly more about the Holmes-style graphs than plain graphs. This included information regarding the subject of the graph, the categories displayed in the graph, the trend of the graph, and the "value message" of the graph. Additionally, participants in the long-term recall group required more prompting to recall graph subject matter, categories, and trends in graphs that did *not* contain embellishments. Participants were also asked to provide preference ratings for the graph styles across a number of factors, and rated Holmes-style graphs significantly higher for more enjoyable, easiest to remember, easiest to remember details, fastest to describe and fastest to remember. The authors note that these findings may be task-dependent, and tasks that require more detailed analysis may be hindered by visual embellishments. They conclude that certain embellishments may improve memory for graph content, and that minimalist design recommendations may not capture the whole picture of graph usability (Bateman et al., 2010). It should also be noted that although their plain graphs certainly have a higher data-ink ratio than the Holmes graphs, they did not evaluate graphs which followed Tufte's recommendations more strictly, and cannot provide further evidence for the "sweet spot" theory. Although this

study did use more complex questions (asking participants to describe the basic trend of a graph), simple and complex questions were not compared directly.

Similarly, Kelly (1989) found no difference in immediate recall of information from high and low data-ink charts in a newspaper format. Using ten bar graphs taken from *USA Today* and ten alternate versions with much of the non-data ink removed, accuracy in making comparisons and in recalling numbers from the graphs was recorded. After a 15-second examination period, participants answered six questions, including "How many bars were displayed in the graph?", "What numerical value was given to the longest bar in the graph?" and "Was the longest bar twice as long or longer than the shortest bar?" For 120 participants, the number of errors made was nearly the same, regardless of graph type. The author concludes that these results provide limited evidence against data-ink maximaization. In the case of both Bateman et al. (2010) and Kelly (1989), it is unclear that recall of descriptive features is an appropriate measure of a graph's effectiveness. However, these results suggest that high data-ink ratios do not result in memorable graphs.

On the other hand, Gillan and Richman (1994) *did* find empirical support for the principle of data-ink maximization. They presented low, medium and high data-ink graphs to 17 undergraduate students, who were asked to answer three types of questions – comparison, difference and mean – using bar and line graphs depicting two data points per graph. They found that the percentage of correct answers was significantly lower for the low data-ink condition than medium and high data-ink conditions. The study's low data-ink graphs were closer to Holmes-style graphs, and used complex background images. There were no performance differences between medium and high data-ink graphs, which followed Tufte's erasing principles more closely, starting from standard-looking bar and line graphs and removing ink. They also found significant differences in response time for all conditions, with high data-ink conditions showing the fastest, low data-ink the slowest and medium in-between. These results suggest support for data-ink maximization in graph design, and use behavioral measures rather than subjective measures. However, there were

only 17 participants (Gillan & Richman, 1994).

In a follow-up experiment, it was found that the presence of a pictorial background in a graph (which lowers data-ink ratio by definition) negatively affected both response time and accuracy. However, it was also found that this effect was more pronounced on difference and mean questions than comparison questions, suggesting an interaction between the presence of a background and question type. The authors conclude that although some of their results support Tufte's suggestions, it appears that the level of data-ink can be either helpful or harmful depending on the task of the graph reader. Overall, they describe finding "support for a more limited approach to graphic minimalism" (Gillan & Richman, 1994, p. 638), which provides further evidence for the idea of a data-ink "sweet spot" that is lower than the maximum.

Blasio and Bisantz (2002) found some support for the principle of data-ink maximization in a simulated monitoring task using dynamic displays which emulated industrial gauges. Twenty-four participants from a diagnostic clinic were recruited to perform a simulated process monitoring task in which nine dynamic variables were to be monitored. Participants monitored either low, medium or high data-ink gauges. Low data-ink gauges included tick marks, number labels, colored regions corresponding to "normal" and "fault" value ranges, and a redundant digital read-out of the "analog" value (shown via a graphic that resembled a vertically-oriented thermometer). Medium data-ink gauges still included the analog representation, but had fewer tick marks, no digital read-out, and no colored regions. High data-ink gauges were simply digital read-outs, with no analog representation of the data. When an "out of range" value appeared on one of the displays, participants were to click a 'Reset' button using a mouse. Participants were significantly faster in identifying out-of-range values using high data-ink displays as opposed to medium or low data-ink displays. However, it should be noted that the high data-ink display was purely numerical (or "digital") while the medium and low data-ink displays were graphical in nature (Blasio & Bisantz, 2002). It could be argued that the

study's high data-ink condition falls outside of the realm of Tufte's principle, as simple numerical values are not graphs. Furthermore, it is difficult to characterize the use of color within the data-ink framework, as color can relay meaning in a way that is independent of the *amount* of ink used.

Finally, Kulla-Mader (2007) found "an overall dislike" of low data-ink graphs and a preference for graphs with medium and high data-ink ratios among 12 participants. Although the study attempted to measure comprehension via questions regarding graph content, level of performance on these questions was too high for response accuracy data to be analyzed, potentially indicating that the questions were simply too easy. It is also interesting to note that the majority of participants described the low, medium and high data-ink graphs as visually similar. Additionally, the low data-ink graph ordered the bars according to magnitude of response (ranging from less than 15 minutes spent using the internet to 4 hours or more) while the medium and high data-ink bars were ordered according to the *number* of responses in each group. Overall, it seems difficult to accept the results of this study as a valid measure of the effects of data-ink ratio.

Although previous research on the data-ink ratio has yielded mixed results, many studies have measured recall of graphed information or design preferences as opposed to performance measures, and have used graphs which do not closely resemble Tufte's high data-ink ratio designs. None of the studies compared performance in graph comprehension tasks using Tufte's "erasing principle" to create graphs similar to his high data-ink re-designs. Furthermore, previous research has failed to account for user and task characteristics which have been identified as important in the process of graph comprehension. For those reasons, it is difficult to accept previous research as a strong test of the data-ink ratio concept, and it seems likely that mixed results are a product of the various methodologies employed in the data-ink ratio literature.

Hypotheses

The previously-described models of graphical comprehension include task demands as an important factor which can interact with characteristics of graph viewers, such as prior knowledge and graph schemas, and with the display characteristics of a graph, such as its data-ink ratio. Zhu (2007) refers to the relationship between visualizations and tasks as the *utility principle*, which states that effective visualizations should aid users in carrying out specific tasks. Much of the past research on the data-ink ratio suggests that the impact of data-ink maximization on comprehension may indeed be task-dependent. Graph viewers use graphs for a variety of reasons, such as understanding causal mechanisms, making comparisons, making decisions, problem-solving and making predictions. Performance in graph comprehension has been assumed to be contingent on congruence between the demands of the viewer's particular task and graphical format. Furthermore, task demands have been shown to interact with other factors, such as graph format, and the usefulness of a particular graph format seems to be dependent on task demands (Shah et al., 2005). It has been shown that changes in the aesthetic features of graphs can have different effects on graph-reading tasks with variable levels of difficulty (Stewart, Cipolla, & Best, 2009). This suggests that if there *is* a benefit to data-ink maximization, it may only be apparent in particular circumstances.

Carpendale (2008) describes a distinction between high-level and low-level tasks in the use of information visualizations. Low-level tasks include compare, contrast, associate, distinguish, rank, cluster, correlate and categorize. More complex, higher-level tasks include understanding trends, uncertainties and causal relationships; making predictions, and learning domains (Carpendale, 2008). An experimental design which better accounts for the varying levels of complexity associated with particular graph reading tasks may provide a clearer picture of the effects of data-ink ratio on graph comprehension.

The present study was conducted in two parts – an experiment which measured response accuracy and mental effort for graphical comprehension questions using varying

levels of both data-ink ratio and task complexity, and semi-structured expert interviews, which provided qualitative feedback regarding the data-ink ratio.

Experiment. Based on the definitions of high and low complexity graph reading tasks described by Carpendale (2008), low task complexity was defined as tasks which require comparing, contrasting, distinguishing or ranking data displayed in a graphs. For example, determining which of two groups in an experiment had the fastest reaction time would be a low complexity task. High task complexity was defined as tasks which require identifying overall trends or understanding causal relationships – for example, making a conclusion about the effectiveness of different training programs.

According to Tufte and supporters of the data-ink hypothesis, increases in data-ink ratio should result in corresponding performance increases in graph comprehension tasks, as measured by response accuracy and mental effort. However, the results of empirical tests of the data-ink ratio, and the frequent suggestion in the literature that the impact of data-ink ratio may be task dependent leads to different predictions. In terms of response accuracy, it is predicted that, overall, accuracy will be higher for low complexity questions than high complexity. It is also predicted that participants will perform similarly well for low complexity tasks, regardless of data-ink level. However, for high complexity questions, it is predicted that accuracy will be lowest in the low data-ink condition (standard graphs), highest in the high data-ink condition (Tufte's redesigned graphs), and somewhere in-between in the medium data-ink condition.

Some have argued that research on information visualizations has defined effectiveness too narrowly by measuring variables such as task completion time, error rate and user satisfaction (Zhu, 2007). Indeed, the majority of published studies on data-ink ratio have used some combination of these measures. However, it has been noted that the speed-accuracy trade off makes it likely that viewers could expend *more* mental effort to achieve the *same* level of performance as measured by response accuracy, making it difficult to judge the overall quality of different graphs and limiting the practical significance of

such findings. These same authors contend that a measure which accounts for the amount of cognitive effort involved in graph comprehension is necessary to evaluate the effectiveness of different graph designs (Huang, Eades, & Hong, 2009).

Mental effort was defined as responses to a self-report Likert scale (Appendix A). It was predicted that high complexity tasks would require greater levels of mental effort than low complexity tasks for each graph type. It was also predicted that low complexity tasks would result in similar levels of mental effort, regardless of the level of data-ink ratio. Finally, it was predicted that for high complexity tasks, increasing data-ink ratio levels would result in decreasing levels of mental effort, with lower overall levels of mental effort for the high complexity/high data-ink condition than the low complexity/high data-ink condition.

Additionally, previous studies on the data-ink ratio concept have failed to account for participant characteristics which can impact graph comprehension, such as the two types of graph schemas described by Pinker (1990). This study collected information about participants' experience with statistics courses, which is assumed to enrich the general graph schema, and experience with using and creating both boxplots and bar graphs, which is assumed to enrich instantiated graph schemas. This allowed performance data to be analyzed in terms of participants' graph knowledge, a potentially important factor in graph comprehension.

Interviews. Learning to interpret graphs has been likened to learning a second language. Those with less practice are assumed to be less capable than experienced graph readers (Carpenter & Shah, 1998). Additionally, the complexity of graph comprehension tasks is assumed to be an important factor in ease of comprehension (Carpendale, 2008), but high-level tasks, such as drawing conclusions or exploring ideas, are difficult to measure using experimental methods (Tory & Moller, 2005). These issues are likely exacerbated when conducting research with undergraduate students, who may have limited experience using graphs. Tufte (2015) has disparaged research on the data-ink ratio concept for using

undergraduate students as participants . That criticism may have merit because models of graph comprehension include graph literacy skills, or *graph schemas*, as an important factor. Therefore qualitative research methods, which focus on detailed narrative information as opposed to statistical inference, can be an important complement to experimental data regarding the data-ink ratio.

Interview techniques have been used to evaluate the user experience of graphs through subjective feedback regarding the effectiveness of different designs in supporting a user's intended tasks. In other words, one goal of interviews is to gather qualitative data which can inform design decisions (Lam, Bertini, Isenberg, Plaisant, & Carpendale, 2012) and result in different conclusions than other research methods (Tory & Moller, 2005). Interview techniques are often used in combination with experimental methods, as they focus on depth of information rather than sample size and allow for data collection in more realistic contexts than traditional experiments (Portigal, 2013). It has been argued that interview techniques should be used more frequently in the evaluation of graphical information (Carpendale, 2008).

It is common to elicit qualitative feedback using a semi-structured interview method (Carpendale, 2008). This usually involves the creation of a document known as a *discussion guide*. This document is not an interview script; rather, it provides necessary structure for the interviewer, such as introductory information, potential interview questions, and a rough outline for the interview. However, it also allows for flexibility during the process (Portigal, 2013). Interviews are usually audio recorded and conducted *in situ*, or in the context in which the relevant activities naturally occur. This brings realism to the interview process, help interviewees to bring to mind examples, and allows reference to work *artifacts* – things that people use, create, modify, or reference in the course of their work (Beyer & Holtzblatt, 1999). In the current study, that could include a journal from the interviewee's field of research or a graph they personally created.

Qualitative interview data can then be analyzed using *thematic analysis*, a flexible

method in which interviewee opinions and interviewer observations are grouped into common themes (Carpendale, 2008). *Themes* represent patterns in responses which relate to the research questions at hand. Researcher judgment is inherent in thematic analysis, and it has been argued that hard-and-fast rules as to what constitutes a theme do not work. However, this method typically involves a number of common steps (Braun & Clarke, 2006). First, recorded verbal data, such as audio recordings of semi-structured interviews, are transcribed. It is important that these transcriptions are both complete and true to the interviews themselves. This stage helps the researcher to develop a clear understanding of the entirety of qualitative data. Next, the researcher reviews the qualitative data to create a list of *codes*, or potentially interesting features about the data. These codes can be further organized into themes, which are then reviewed. The purpose of review is to ensure that the contained data within each theme is internally consistent, and themes are meaningfully distinct. Finally, the resulting themes are defined and named (Braun & Clarke, 2006).

Experiment Method

Participants

A convenience sample of 175 undergraduate students (53 women, 122 men, M_{age}=19.8, age range: 18-46 years) enrolled in three different sections of an introductory psychology course at RIT was recruited for participation. Although the sample is roughly 70% men, that is similar to the percentage of men in the RIT undergraduate population (*RIT in Brief*, 2015). Data collection sessions took place during the courses' regularly scheduled meeting times. Data were collected from 179 students, but incomplete data from four students were not included in analysis. Participation in the experiment partially fulfilled the research participation requirement for the course – participants received two participation credits on the SONA research participation system.

Roughly half the sample comprised first-year students (N=87). An additional 52 students were in their second year of studies, and the remaining 36 students were in their

third year or higher. Participants came from a variety of academic majors, including computer science, game design and development, and various sub-disciplines of engineering. Seventy-one percent of the sample majored in science/engineering (N=125), 22% majored in arts/humanities (N=39), and 6% majored in social sciences (N=11). Although data collection took place exclusively in psychology courses, none of the sample reported psychology as their major. Seventy-four participants reported having taken a statistics class, as opposed to 101 who had not.

In general, the sample was experienced with bar graphs. Ninety-nine percent (N=173) of the sample reported having seen a bar graph prior to participation, 97% (N=170) reported having used a bar graph prior to participation, and 97% (N=169) reported having made a bar graph prior to participation. Boxplot experience was more variable – roughly 80% (N=138) of the sample reported having seen a boxplot prior to participation. About 63% (N=111) of the sample reported having used a boxplot prior to participation. Finally, 52% (N=91) of the sample reported having made a boxplot prior to participation, as opposed to 48% (N=84) who had not. Experience with statistics courses was related to boxplot experience – all participants who had taken a statistics course had seen a boxplot prior to participation, and nearly all had used and created a boxplot. Of the 101 participants who had *not* taken a statistics class, smaller numbers had experience with boxplots (Table 1).

Table 1

Boxplot experience by experience with a statistics class

	Seen boxplot	Used boxplot	Made boxplot
Statistics class (N = 74)	74	70	68
No statistics class (N = 101)	64	41	23

Materials

The experiment used a 3x2x2 mixed design with two levels of task complexity (high and low) and two graph types (bar and boxplot) as within-subjects variables and three levels of data-ink ratio (high, medium, and low) as a between-subjects variable. Bar graphs and boxplots were chosen because Tufte provides examples of high data-ink versions of those graphs. Individual participants answered a common set of questions using one bar graph and one boxplot with matched levels of data-ink ratio.

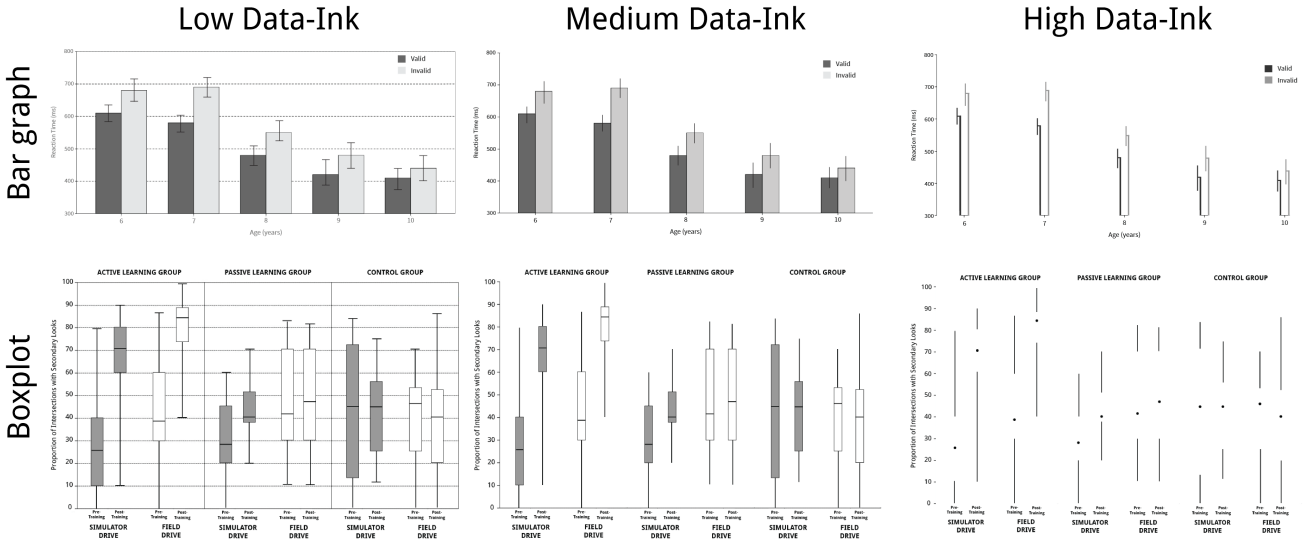


Figure 7. Experimental stimuli: low, medium, and high data-ink bar graph and boxplots.

Using Adobe Illustrator v1.6, two types of graphs were adapted from published psychology studies – bar graphs (Lellis et al., 2013) and boxplots (Romoser & Fisher, 2009). Three versions of each graph were created, corresponding to three data-ink ratio levels – low, medium, and high (Figure 7). Low data-ink graphs were generally similar to the published originals, but included additional features which Tufte specifically recommends against, such as dashed gridlines, vertical separating lines between grouped graphical elements, and thicker bars and boxes. High data-ink graph designs mimic Tufte's (2001) high data-ink redesigns of bar graphs and boxplots, as shown in Figures 2 and 3. By placing different graph elements in different image layers, it was possible to follow Tufte's

"erasing principles" by simply hiding particular layers, such as gridlines or the "box" portion of a boxplot. Although the size of some graph elements was reduced, their absolute locations on the graph were not changed. The overall size of the graphs was held constant. The medium data-ink graphs served as an intermediate between high and low data-ink graphs – gridlines, T-intersections on bar graph error bars and boxplot whiskers, and non-axis border lines were removed. Additionally, the thickness of box-shaped visual elements was reduced by one third relative to the low data-ink versions. All graphs were created in grayscale.

To accompany the graphs and provide context, written explanations were created to describe the experiments from which the graphs were taken. These explanations described the methods and stimuli of the experiment and defined relevant terminology as necessary. They did not, however, describe the results of the experiment or any conclusions reached by experimenters. An example graph and accompanying explanation are shown in Figure 8, and both graph explanations are shown in Appendix B.

Graph Lesson. To account for varying levels of experience with graph comprehension tasks and the graph types used in the study, participants were given a two to three minute presentation on the features of bar graphs and boxplots immediately before beginning the experiment. This included information regarding axes, bar length, error bars, quartiles, medians, whiskers, and maximum and minimum values. The accompanying PowerPoint presentation showed example graphs which were similar in design to the medium data-ink ratio graphs used in the experiment. As each individual graph feature was discussed, it was highlighted in the presentation using red circles (Appendix C).

Graph comprehension questions. Response accuracy was measured using open-ended questions regarding the content of the graphs. The questions were developed to correspond to the previously-described low and high complexity task definitions – low complexity questions required participants to compare, contrast, or rank while high complexity questions required participants to understand trends and causal relationships or

The following graph depicts the results of an experiment which measured *secondary looks* – the act of glancing in the most likely direction of oncoming traffic after beginning a turn – in a group of older drivers. All participants completed pre-training simulator and field tests in which the proportion of intersections with secondary looks was measured. Next, one group received active, immersive training using the simulator while a second group received passive, classroom-style training. A control group received no further training. All 3 groups were then tested again in both the simulator and in the field, with the proportion of intersections with secondary looks measured again.

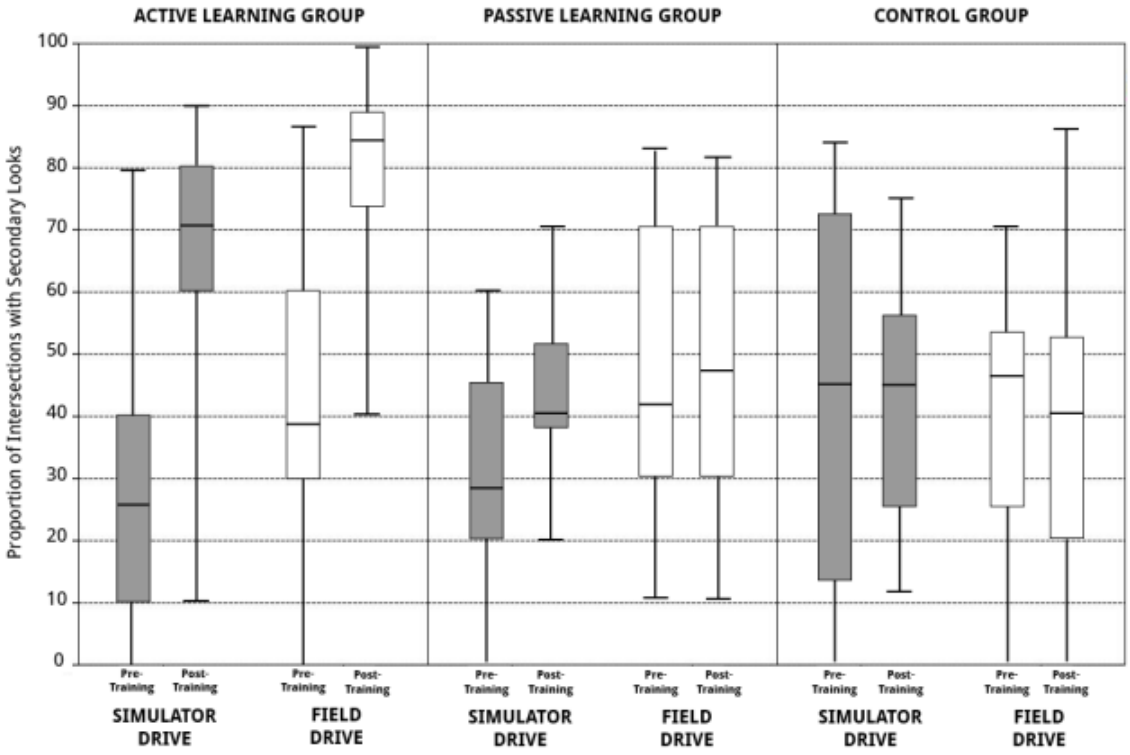


Figure 1. Pre- to post-training changes in proportion of intersections with secondary looks for both simulator and field drives. Plot whiskers represent minimum and maximum proportions observed.

Figure 8. Example experimental stimuli for low data-ink ratio boxplot.

make predictions. Two low complexity questions and two high complexity questions were created for each graph, resulting in a total of eight questions per participant (Appendix D).

Mental Effort. Mental effort was measured using an adapted version of the Paas

scale, a 9-item Likert scale ranging from "very, very low mental effort" to "very, very high mental effort." This scale was shown to be reliable in a study which tested problem-solving skills and cognitive load in statistics (Paas, 1992) and has been used successfully in studies exploring the effectiveness of graphs (Huang et al., 2009). However, the wording of some items was changed for greater consistency within the scale (Appendix A).

Demographics questionnaire. After answering graph questions, participants also completed a demographic questionnaire. In addition to age and gender information, participants reported their academic major, information regarding their experience with statistics courses, and a series of questions regarding their previous experience with both bar graphs and boxplots (Appendix E).

Procedure

Each data collection session began with the distribution of packets which included the necessary materials for participation. After an explanation of the purpose of the study and time to read and sign a consent form, students were given the graph lesson and the experiment began. Because the graphs were presented in paper format, participants were free to refer to the graph at any point during the experiment and refer to the questions and graph descriptions as necessary. Participants were instructed to remove the page containing the graph if they wished to do so, as the graph and questions spanned several pages.

To account for order effects, the sequence of graph presentation was balanced across packets – for roughly half of participants, the bar graph appeared first and for the other half the boxplot appeared first. Additionally, participants answered questions in one of two randomly generated orders. This resulted in two versions for each data-ink level, and six versions in total. However, given the nature of the data collection method, participants could have answered questions in any order. After answering the eight graph questions, participants completed the demographic questionnaire. Participants were not given a time limit. The majority of participants completed the task in 12-13 minutes. Consent forms and data were collected separately.

Packet versions were sorted prior to data collection to ensure even distribution throughout the sample. In total, 58 participants used low data-ink graphs (30 version 1, 28 version 2), 61 used medium (31 version 1, 30 version 2), and 59 used high (30 version 1, 29 version 2).

Prior to data collection, two pilot tests were conducted. After an initial round of pilot testing, changes were made to the graph lesson, the wording of individual questions, and to the structure of the experimental "packet". After an additional round of pilot testing, the initial set of 12 questions (six per graph) was reduced to a set of eight (four per graph). Those eight questions were selected based on the distribution of pilot results – questions with normally distributed responses were favored over questions with skewed results and/or multiple outliers.

Interview Method

Participants

Seven interviews were conducted with faculty members from the Rochester Institute of Technology (RIT) with a variety of academic backgrounds. Five of seven interviewees held doctorate degrees. Three of those were in psychology, one was in psychophysiology and one was in industrial engineering, but taught courses in applied statistics. The other two interviewees held a master's degree – one in graphic design and the other an MFA in visual and verbal communication (the terminal degree in their field). A strict selection criteria was not used, but preference was given to those who were likely to have opinions regarding graph design (e.g., faculty in design, human factors and statistics) and/or those with frequent graph use. Participants were found through recommendations by other faculty members or through departmental web pages.

Procedure

Prior to the interviews, a pilot interview was conducted to develop an effective interview technique that would elicit pertinent information. Interviewees were recruited from the faculty at the Rochester Institute of Technology (RIT). After agreeing to

participate, interviewees were sent a common set of nine pre-interview questions via e-mail (see Appendix F). These questions were general in nature, focusing on graph use and creation, and participants' responses were used to create a discussion guides tailored to the interviewee (see Appendix G for example). Two interviewees had prior knowledge of the study and interview methodology, but it was determined that their responses were not fundamentally different from other interviewees prior to inclusion.

Each interview lasted roughly one hour and focused on the use and creation of graphs, context of graph use, the importance of aesthetics in graph design, knowledge of and opinions about the data-ink ratio concept, and feedback on example graphs with varying data-ink ratios. The example graphs were the same bar graphs and boxplots that were used in the experiment. Interviews were conducted in participants' offices to allow access to personal materials, research publications, graph-making software, or any other work artifact that the interviewee wished to reference. Audio recordings of the interviews (recorded with a Sony ICD-PX312) were summarized and synthesized using thematic analysis. Interviewees were given a gift certificate ($10 value) for their participation in the interview, but were not aware of any remuneration at the time they agreed to participate. Gift certificate funding was provided by RIT's College of Liberal Arts.

Experiment Results

Bar graph. On average, participants responded to bar graph questions with 79% accuracy, *SD*=.22, 95% CI [.75, .82]. Mean accuracy for low complexity bar graph questions was 84%, *SD*=.27, 95% CI [.80, .88], which was significantly higher than a 74% mean accuracy for high complexity bar graph questions, *SD*=.31, 95% CI [.69, .79]. The mean mental effort rating for all bar graph questions was 3.1, *SD*=1.1, 95% CI [2.9, 3.3]. The mean mental effort rating for low complexity questions was 2.6, *SD*=1.2, 95% CI [2.4, 2.8], which was significantly lower than the mean mental effort rating of 3.6 for high complexity bar graph questions, *SD*=1.2, 95% CI [3.4, 3.8].

There was no effect of data-ink level on mean accuracy for all bar graph questions

(both low and high complexity). However, there was one interactive effect of complexity and data-ink level on accuracy – participants answered low complexity/high data-ink bar graph questions (M=.87, SD= .24, 95% CIs [.81, .93]) significantly more accurately than high complexity/high data-ink bar graph questions (M=.70, SD= .32, 95% CIs [.61, .78]). High complexity questions generally resulted in slightly higher mental effort ratings than low complexity questions, but there was no effect of data-ink level on mental effort ratings for all bar graph questions, for low complexity bar graph questions, or for high complexity bar graph questions (Figure 9).

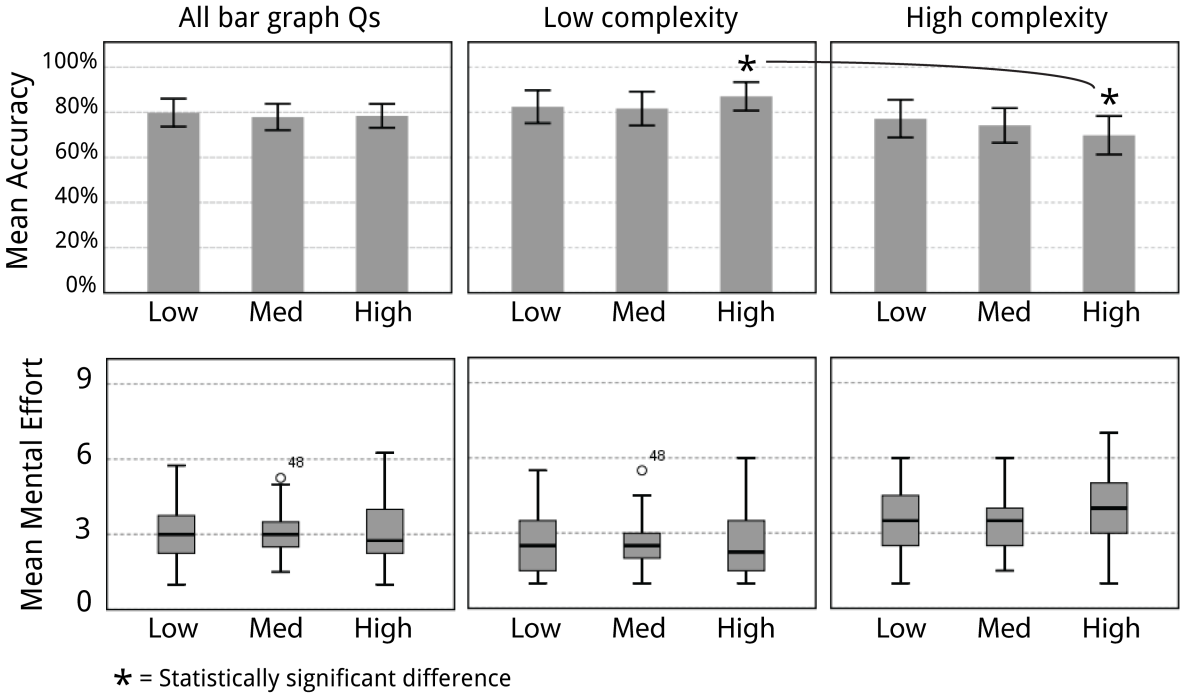


Figure 9. Mean accuracy and mental effort ratings for all bar graph questions, low complexity bar graph questions, and high complexity bar graph questions. Data-ink levels are displayed along the x-axes. Error bars represent 95% confidence intervals. Boxplot whiskers represent 1.5 x inter-quartile range.

Experience with a statistics course did not have an interactive effect with data-ink level on accuracy – although participants who had taken a statistics course performed slightly more accurately in all conditions than participants who had not taken a statistics

course, none of those differences were significant. Regardless of experience with a statistics course, mental effort ratings were similar across conditions (Figure 10). Additionally, year of study did not have any interactive effects with data-ink level or complexity. There were no statistically significant differences in accuracy between participants in their first year of study and participants in their second year of study or higher (Figure 11).

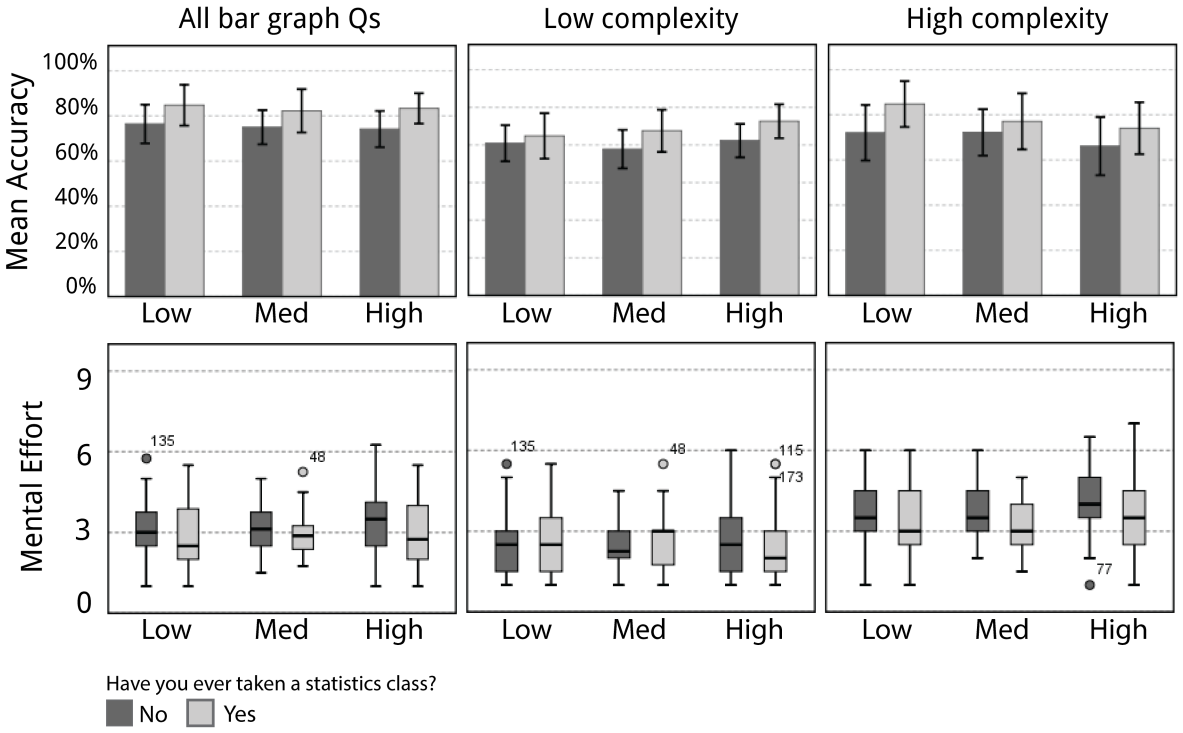


Figure 10. Mean accuracy and mental effort ratings for all bar graph questions, low complexity bar graph questions, and high complexity bar graph questions. Results are further broken down by previous experience with a statistics course. Data-ink levels are displayed along the x-axes. Error bars represent 95% confidence intervals. Boxplot whiskers represent 1.5 x inter-quartile range.

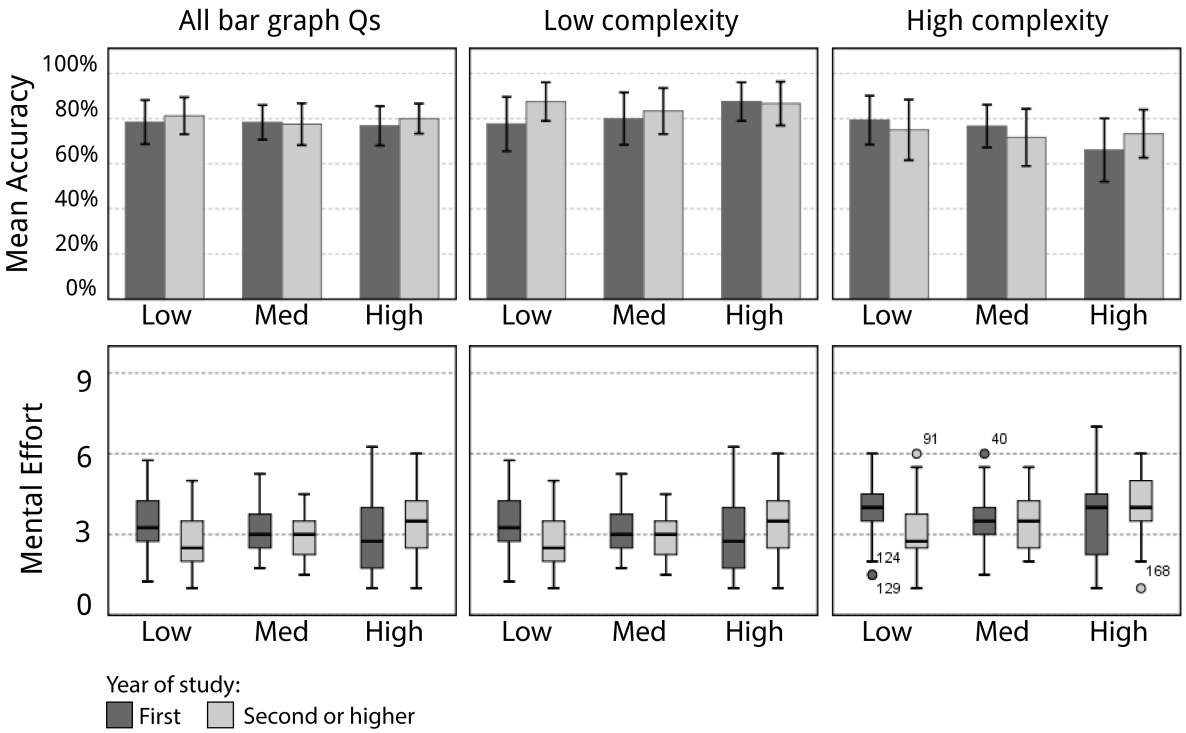


Figure 11. Mean accuracy and mental effort ratings for all bar graph questions, low complexity bar graph questions, and high complexity bar graph questions. Results are further broken down by year of study. Data-ink levels are displayed along the x-axes. Error bars represent 95% confidence intervals. Boxplot whiskers represent 1.5 x inter-quartile range.

Boxplot. Boxplot questions were answered with 55% accuracy, *SD*=.32, 95% CI [.50, .60], which was significantly lower than overall bar graph accuracy. Participants responded slightly more accurately to high complexity questions (*M*=57%, *SD*=.36, 95% CI [.51, .62]) than low complexity questions (*M*=53%, *SD*=.40, 95% CI [.47, .59]), but that difference was not statistically significant. However, the mean mental effort rating of 3.6, *SD*=1.3, 95% CI [3.4, 3.8] for high complexity questions was significantly higher than the mean mental effort rating for low complexity questions (*M*=3, *SD*=1.3, 95% CI [2.8, 3.2]). The overall mean mental effort rating for boxplot questions was 3.3, *SD*=1.2, 95% CI [3.1, 3.5], which was not significantly different than the mean mental effort rating for bar graph questions.

As with bar graph questions, there were no effects of data-ink level on mean accuracy or on mental effort ratings. This was the case for all boxplot questions (both high and low complexity), for low complexity boxplot questions only, and for high complexity boxplot questions only. Participants performed least accurately using high data-ink graphs in both the low and high complexity conditions, but those differences were not statistically significant. For low complexity questions, mental effort ratings were slightly lower in the medium data-ink condition than low and high data-ink conditions (Figure 12).

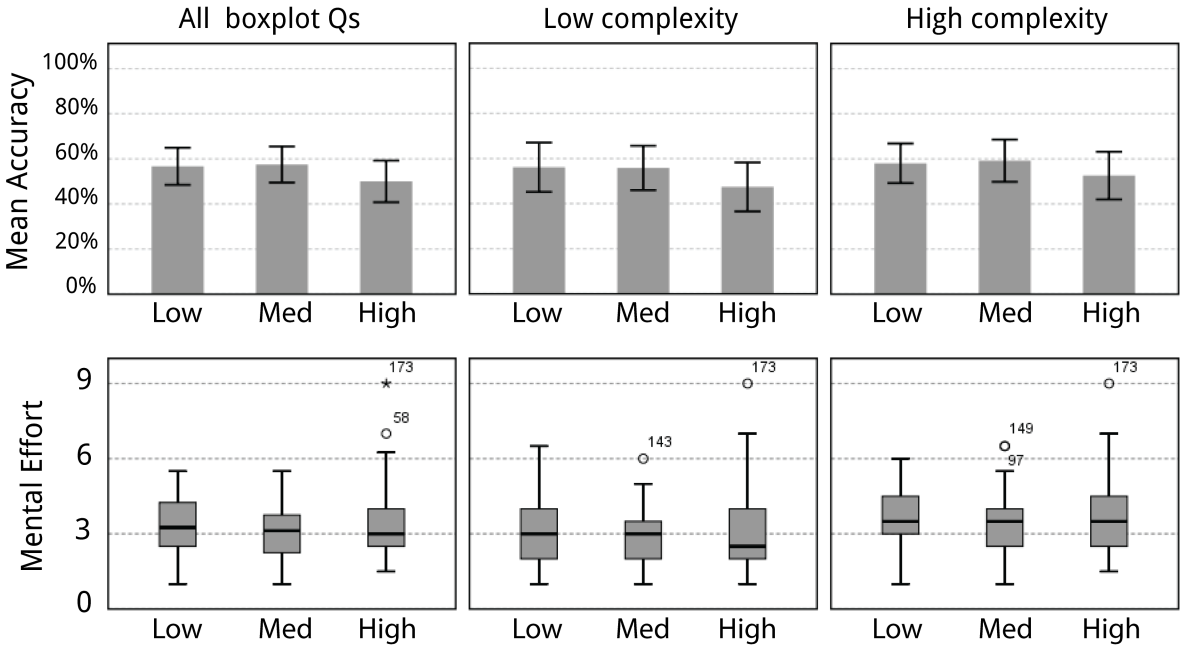


Figure 12. Mean accuracy and mental effort ratings for all boxplot questions, low complexity boxplot questions, and high complexity boxplot questions. Data-ink levels are displayed along the x-axes. Error bars represent 95% confidence intervals. Boxplot whiskers represent 1.5 x inter-quartile range.

Previous experience with a statistics course did not have any interactive effects with data-ink level or with complexity level. In overall accuracy (both high and low complexity questions), participants who had taken a statistics course performed more accurately on average than participants who had not taken a statistics course in the low and medium

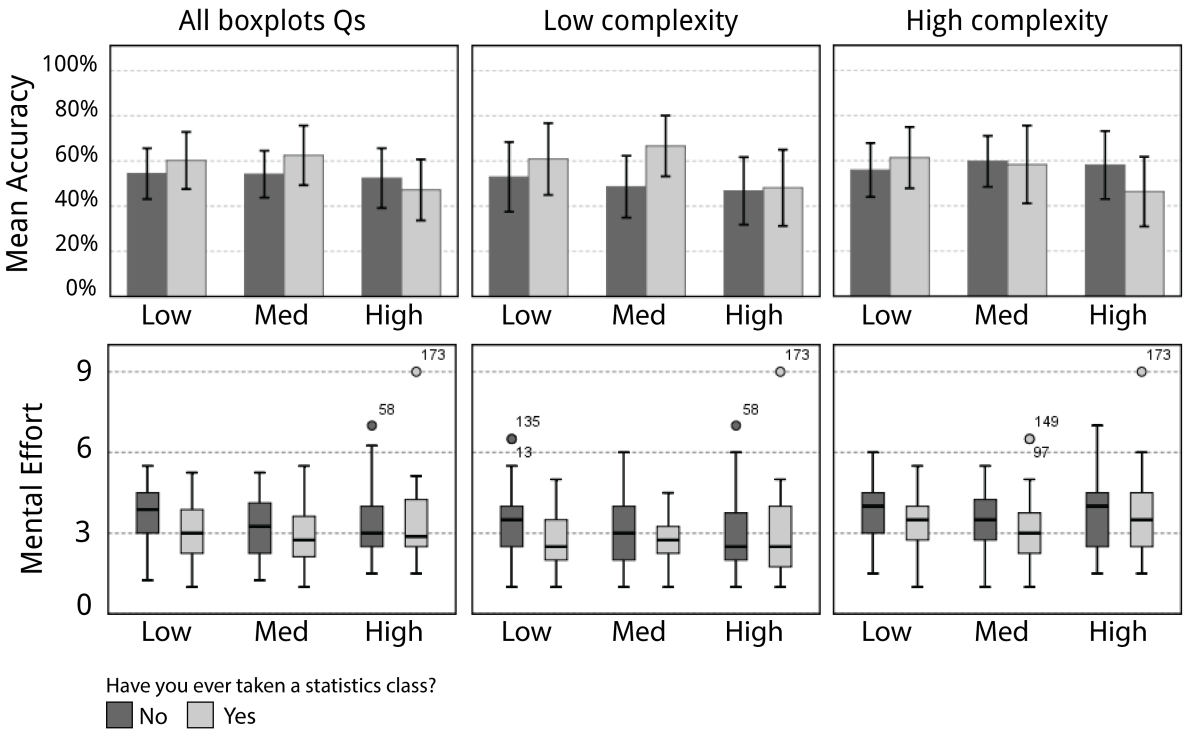


Figure 13. Mean accuracy and mental effort ratings for all boxplot questions, low complexity boxplot questions, and high complexity boxplot questions. Results are further broken down by previous experience with a statistics course. Data-ink levels are displayed along the x-axes. Error bars represent 95% confidence intervals. Boxplot whiskers represent 1.5 x inter-quartile range.

data-ink groups, but *lower* in the high data-ink group. However, none of those accuracy differences were statistically significant. Mean accuracy scores were more similar between participants with and without experience with a statistics course in the high complexity condition than in the low complexity condition. However, those differences were also not statistically significant (Figure 13). As with bar graph questions, year of study did not have interactive effects on accuracy with boxplot questions (Figure 14).

Previous experience with boxplot creation also did not have any interactive effects with data-ink level or complexity level on accuracy. Interestingly, participants who reported previous experience with boxplot creation (91 participants, as opposed to 84 who had not) had *lower* mean accuracy scores in all but one condition – low

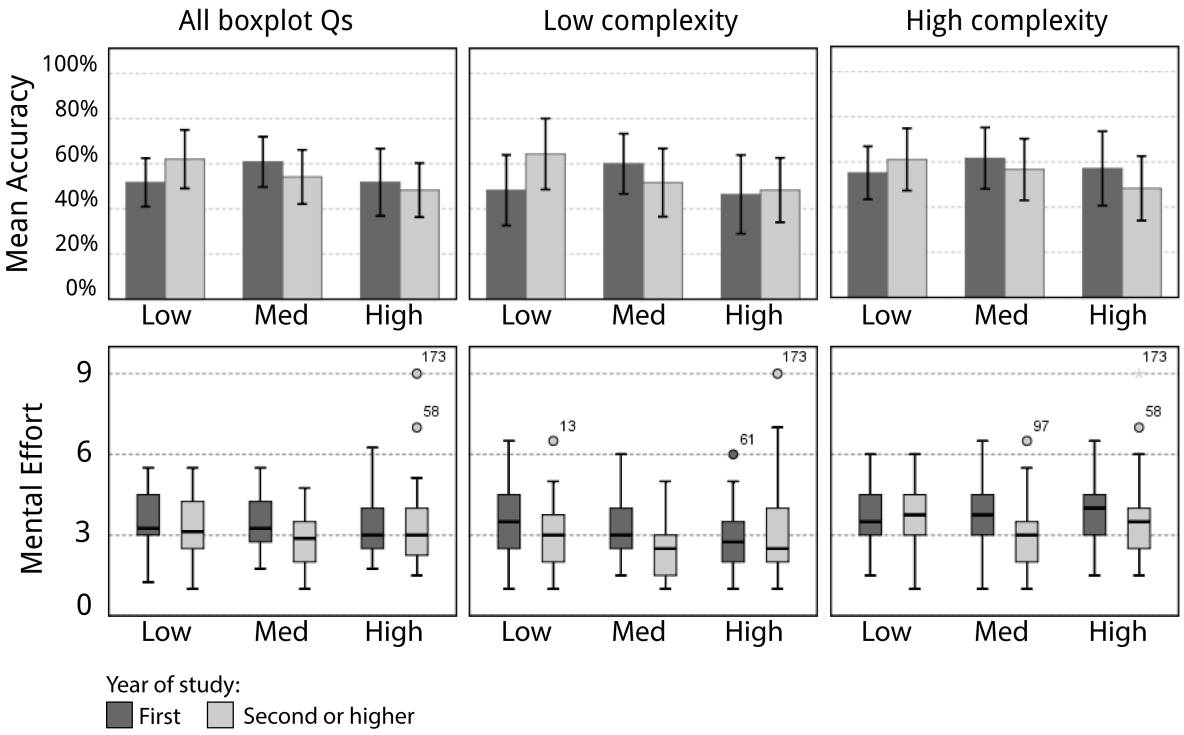


Figure 14. Mean accuracy and mental effort ratings for all boxplot questions, low complexity boxplot questions, and high complexity boxlot questions. Results are further broken down by year of study. Data-ink levels are displayed along the x-axes. Error bars represent 95% confidence intervals. Boxplot whiskers represent 1.5 x inter-quartile range.

complexity/medium data-ink ratio. However, none of those differences were statistically significant. In most cases, participants who reported previous experience with boxplot creation had slightly lower mental effort ratings. However, overall mental effort ratings using high data-ink graphs were more variable among participants with previous experience with boxplots (Figure 15). For complete tables of experiment results, including accuracy and mental effort data for individual graph comprehension questions, see Appendix H.

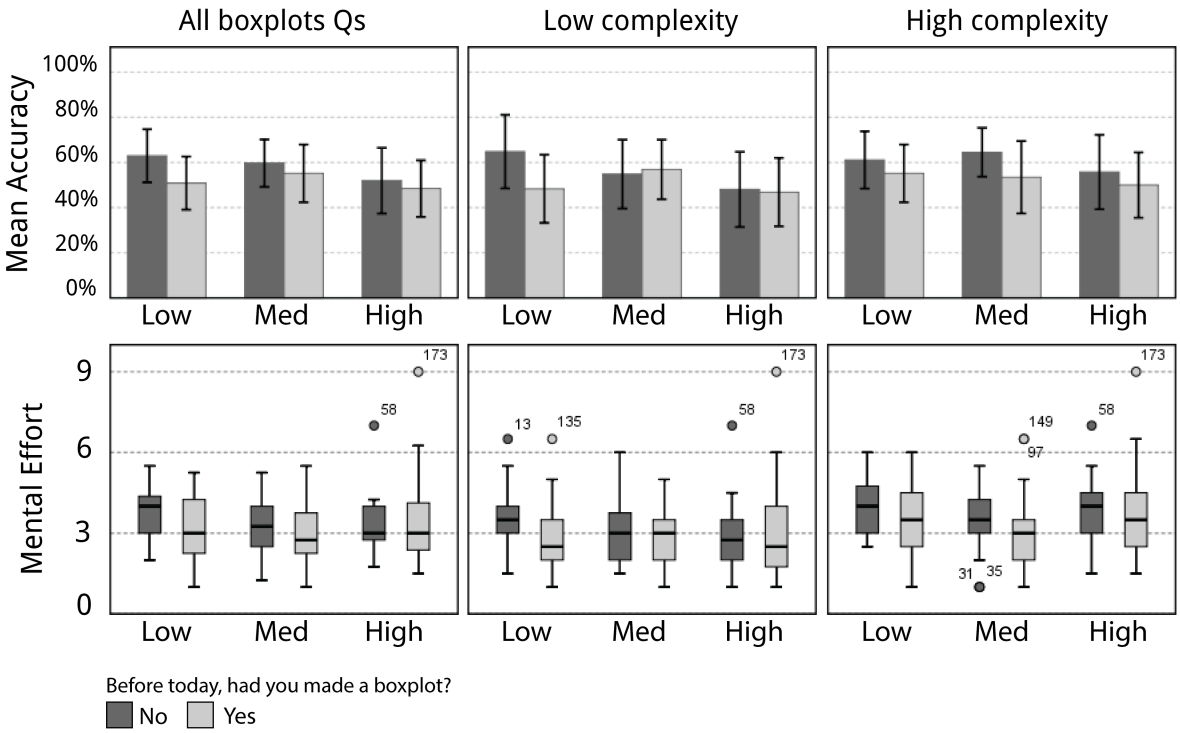


Figure 15. Mean accuracy and mental effort ratings for all boxplot questions, low complexity boxplot questions, and high complexity boxplot questions. Results are further broken down by previous experience boxplot creation. Data-ink levels are displayed along the x-axes. Error bars represent 95% confidence intervals. Boxplot whiskers represent 1.5 x inter-quartile range.

Mental effort reanalysis. In addition to the aforementioned analyses, mental effort ratings for both bar graph and boxplot questions were reanalyzed in two ways. First, the scale was collapsed using its semantic features. Ratings of 1-4 (all variations of "low mental effort") were recoded as a rating of 1, ratings of 5 ("neither high nor now mental effort") were recoded as a rating of 2, and ratings of 6-9 (all variation of "high mental effort") were recoded as a rating of 3. In a second re-analysis, mental effort ratings were collapsed numerically. Ratings of 1-3 were re-coded as a rating of 1, ratings of 4-6 were recoded as a 2, and ratings of 7-9 were re-coded as a rating of 3. In both cases, collapsing mental effort scores failed to reveal any effects.

Unexpected graph annotations. Fourteen participants returned graphs with additional features drawn on them (see Appendix I for complete set of images). Eleven high data-ink graphs were annotated and four medium data-ink graphs were annotated. No features were found added to low data-ink graphs. However, some participants did not return detached graph pages with their data packets, so this list cannot be considered exhaustive.

Five participants added bar graph features to high data-ink graphs and 3 to medium data-ink graphs. Three of 8 of these participants reported having taken a statistics class, and 6 of 8 reported having created a bar graph prior to participation. All participants who modified bar graphs reported having seen a bar graph graph prior to participation. Six participants added features to high-data ink boxplots and 1 added features to a medium data-ink boxplot. Five of 7 boxplot modifiers reporting having taken a statistics class and 6 of 7 reported having created a boxplot prior to participation.

Bar graph annotations included horizontal lines drawn from bars to the y-axis or from on bar to another (Figure 16), vertical separating lines between grouped boxplot elements (Figure 17), and boxes in place of white space in high data-ink boxplots (Figure 18). Six participants added features to a bar graph, 6 added features to a boxplot, and 2 added features to both. Added bar graph features included horizontal lines drawn across to the y-axis or to other graph elements (6 participants), additional tick marks on the y-axis (1 participant) and numerical labels on particular bars (1 participant). Boxplot annotations included vertical separating lines between grouped elements (2 participants), boxes in place of white space in high-data ink graphs (2 participants), and horizontal tick marks, either at the ends of vertical line elements or median dots or to the y-axis (3 participants). For a complete set of annotations, see Appendix I.

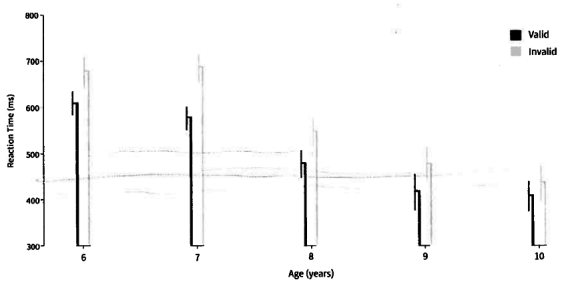


Figure 16. Example of a participant who drew horizontal lines on a high data-ink bar graph. Brightness and contrast adjustments were made to improve annotation visibility.

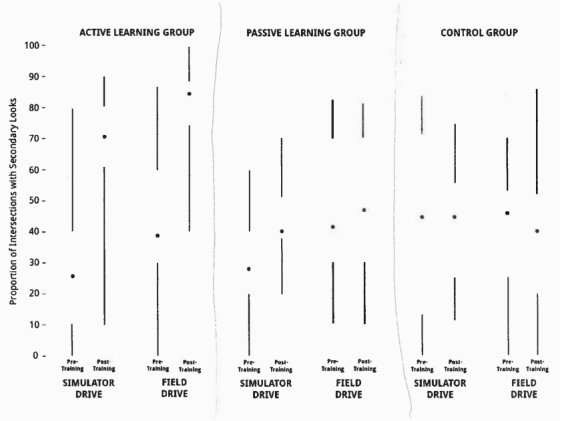


Figure 17. Example of a participant who drew vertical lines on a high data-ink boxplot. Brightness and contrast adjustments were made to improve annotation visibility.

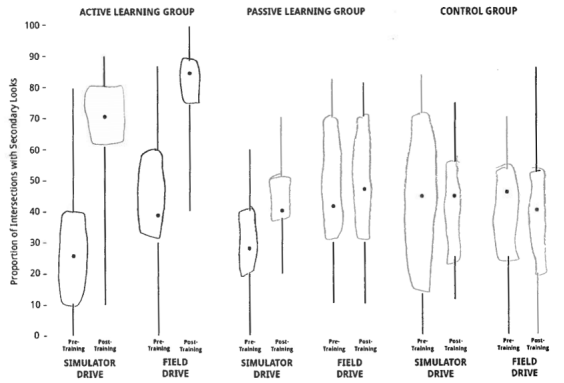


Figure 18. Example of a participant who drew boxes on a high data-ink boxplot.

Interview Results

Data-ink ratio and example graphs. Three interviewees were familiar with the data-ink ratio concept and provided opinions about it. One of those three had a background in design and the others had backgrounds in psychology and imagining science. Two of those interviewees had personal copies of Tufte's book. One of the three described the data-ink ratio as a "neat idea" and agreed that graph features with no relevance should be removed. However, like Carswell (1992), that interviewee also expressed doubts as to whether data-ink ratios can actually be measured and did not believe that the data-ink ratio should be *maximized*, but rather that there is a "sweet spot" for data-ink levels which is lower than the maximum. This was because making graph elements too small makes them difficult to see and additional elements, such as gridlines, can be helpful to guide viewers' eyes. This interviewee reported that he did not apply the data-ink ratio to the design of graphs he creates. The interviewee with an imaging science background described the data-ink concept as a design argument that didn't result in more usable graphs. Finally, a third interviewee, who had a background in design, felt much more positively about the data-ink concept and followed and taught many of Tufte's recommendations for graph creation. The remaining four interviewees were either unfamiliar with the data-ink concept or only vaguely familiar with it.

Feedback regarding the low data-ink bar graph tended to be negative or neutral. It was described as both "fat" and "chunky" by different interviewees, suggesting that the bars were seen as disproportionately large for the size of the graph. One interviewee described it as heavy handed, not due to the size of the bars, but because of the "noise" in the form of gridlines, tick marks, and other elements that could be described as non-data-ink. A different interviewee disliked the box around the graph. On the other hand, the graph was also described as having "some nice elements" – the T-intersections on the error bars were seen as helpful and the gridlines were not "too heavy," but could have been fainter. Another interviewee identified this as their favorite bar graph version, as the

gridlines were helpful due to width of the graph. That interviewee also found T-intersections at the end of error bars to be helpful.

One participant described the medium data-ink graph as "more pleasing" than the low data-ink bar graph due to the increased white space and thinner bars, but would have added faint gridlines and T-intersections to the error bars. On the other hand, a different participant felt that the bars should have been closer together to facilitate comparisons, but identified the graph as their favorite bar graph version nonetheless.

Two participants felt that the high data-ink bar graph would take longer to interpret than the other versions, although one did note that familiarity with the high data-ink style might make it easier to use. An additional interviewee described the graph as "horrible." On the other hand, a different interviewee found this graph to be elegant and minimal, but unnecessarily wide due to the increase in white space created by the thin bars. That interviewee also noted that adding gridlines to this graph would have created a criss-cross pattern with the thin bars. When asked whether moving the bar pairs closer together would improve this graph, a different interviewee said that it would, but that such a change would not impact her negative feelings toward the graph. One interviewee felt that there was "less in the way" in the high data-ink bar graph, and that it could be improved further by removing the "bar" portions of the graph. This interviewee saw bars in general as a waste of ink which might not add anything, as the error bars are the key information. Another interviewee felt that the high data-ink bar graph had been "cleaned up" compared to the others, but that the bars could be thicker to make it easier to differentiate between their colors. This is similar to to Few's (2009) argument that graph elements can be over-reduced.

The low data-ink boxplot was generally described as too busy. More interviewees gave negative comments about the gridlines in this graph than about the bar graph gridlines. Although they were the same size and color as the gridlines in the bar graph, there were a greater number in the boxplot (4 vs. 9, respectively), suggesting that opinions

regarding the inclusion of gridlines are dependent upon the specific graph. One interviewee felt that the T-intersections at the ends of the whisker portions were unnecessary.

The medium data-ink boxplot received more positive feedback than the low data-ink boxplot, though many interviewees suggested changes to the design which they felt would improve it. One interviewee commented that T-intersections at the ends of whiskers and subtle gridlines would improve the graph. Another would have liked vertical separating lines between the experimental conditions. A third interviewee said that a hybrid of the low and medium data-ink graphs would be ideal, though a specific combination of features was not mentioned.

The high data-ink boxplot was widely disliked – all but one interviewee found it hard to read. It was noted that the box portion, present in the low and medium data-ink boxplots, helps to make each individual boxplot look like a cohesive unit. This is similar to Kosslyn's (1985) argument that completing forms results in fewer perceptual units. One interviewee commented that the graph required too many "mental gymnastics," and wasn't sure that she would have known it was a boxplot in a different context. On the other hand, a different interviewee felt that the high data-ink boxplot "says the same thing as the others," but does so more efficiently. Additionally, that interviewee felt that the high data-ink design would be accepted with time, and that the other designs may eventually look archaic. Finally, two interviewees who gave negative feedback about this graph commented that it *does* highlight the trend of median values in the graph due to the large amount of white space around them.

Graph use. Interviewees reported using graphs for a variety of reasons, including publishing empirical results, understanding others' research, teaching courses, measuring student progress in courses, evaluating the effectiveness of interventions, and more. Frequency of graph use ranged from daily, to weekly, to multiple times over the course of a semester. Two interviewees described their graph use as "very frequent" and "extremely frequent," respectively. Heavy graph use was reported when involved in research projects.

Bar graphs, scatterplots and line graphs were commonly encountered among interviewees, though other types, such as radial graphs, boxplots, ISOTYPE and histograms were also mentioned. Some interviewees preferred to use particular types of graphs, such as bar graphs, because of ease of interpretation, or boxplots because they show complete distributions. One interviewee had a preference for graphs that plotted every data point. Others didn't have preferences for particular types of graphs, and instead preferred whichever graph was most appropriate for the particular situation.

Graph creation. Interviewees created graphs using a variety of software tools, including Excel, SPSS, R Statistics, Adobe Illustrator, InDesign, MATLAB, and Jmp. Some interviewees used multiple programs for graph creation, choosing whichever is more appropriate (or easier) for a given graph creation task. For example, R Statistics was described as allowing the most customization of graphs, which was not always seen as necessary for all occasions. Two interviewees reported sketching graphs by hand when early in the graph design process, which was described as a way to avoid the limitations of software and find the best way to display the data.

A number of salient themes emerged on the topic of graph creation goals. Nearly all interviewees named *clarity* as a design goal, which was defined as readability or "ease of use," as well as avoiding clutter. Interviewees wanted their graphs to be understood by others with little effort. One interviewee noted that context is important for clarity. For example, she expected that any graph would be described in writing prior to appearing in a research article to provide a context for that graph. *Accuracy* was also mentioned frequently as a design goal – graphs should show the data as they actually are without obscuring phenomena. The use of truncated axes was the typical example of inaccuracy or dishonesty in graph design, and multiple interviewees noted that accuracy problems are frequently unintentional, but problematic nonetheless. One interviewee identified accuracy as the most important factor in graph design, and noted that it can be difficult to judge the accuracy of graphs created by others.

All interviewees were conscious of the aesthetics of graphs they create, but had a variety of definitions for this concept. Some used words like “clean” or “elegant” to describe their goals with regard to aesthetics. Features such as the size, color, and placement of graph elements were seen to impact the aesthetics of a graph. Both of the interviewees with a design background mentioned “balance” as a graph design goal – the idea that a graph creator must make trade-offs between simplicity, visual interest, clarity, and completeness. This design goal was reminiscent of Tukey’s advice for “well-tuned” graphs. Some interviewees described graph-making conventions as “heavy-handed” or even ugly, and nearly all interviewees expressed some level of dissatisfaction with the look of default designs created by graph-making software.

Effective labeling was critical to a number of interviewees – three reported that labels are among the first features of graphs that they read, and that they are helpful for identifying the variables or conditions in an experiment. Good labeling was said to include titles, legends, and figure captions. One interviewee felt that good labeling is more important than limiting visual clutter. A different participant mentioned the importance of the layout of labels in graph design – for example, labels shouldn’t be shown at 90 degree angles and the alignment of labels should be consistent for improved readability.

Gestalt principles were mentioned in multiple interviews by those with both psychology and design backgrounds. A designer noted that gestalt principles are foundational to design education. Features such as color and grouping via proximity were seen as important to good graph design. Symbols were also used to group graphed data. The principle of closure was implicitly discussed – one interviewee noted that the “box” portion of a boxplot helps each element to look like a cohesive unit.

The importance of matching graph type to data type was emphasized by three interviewees. For example, bar charts were said to be appropriate for comparing categorical data, while scatterplots or line graphs would be appropriate for trend data. Both of these interviewees had seen graphs which were *not* appropriate for the data they showed. This

was seen as an aspect of graph design that requires particular skill and knowledge. The match between graph types and data types was described as a type of "natural mapping," which was also important to other factors of graph design, such as matching the orientation of a legend to the orientation of the graphs features it represents. One of these interviewees stressed that graph creators should not create only *one* type of graph for a given situation, but that graph types *can* be used inappropriately.

Hierarchical structure in graph design was mentioned by two interviewees. One reported that the data should always be primary in visual emphasis, that features such as axes and legends should be secondary, and that gridlines and tick marks should be tertiary. Primary content should "come up to the surface" while everything else "recedes to a place of secondary of tertiary importance." If anything "upstages" the data, it is a design problem which can make a graph "heavy-handed." The other interviewee reported that the "most important things" in a graph should be emphasized in the design, and that the designer should know what the hierarchy of their graph is. For example, if a line graph is being used to show trend data, the line portion is most important, and that element should be bolder than elements such as axes or tick marks.

Interviewees had few absolute rules with regard to graph creation – the majority of design choices described during the interviews were dependent upon the specific features of the data and context of presentation. Interviewees did not want graphs to be "busy" or include superfluous features, but definitions of what constitutes *superfluous* varied between participants and situations. For example, some interviewees reported disliking gridlines in graphs, but mentioned scenarios in which they might include them, such as allowing readers to track across a particularly wide graph or the use of a single gridline to highlight a particular value. Similarly, two interviewees noted that they might include gridlines for their own use as a measurement tool during data exploration, but remove them when creating graphs for the public. One interviewee liked Tufte's suggestion to create gridlines using thin white lines on the bars themselves. Similarly, some interviewees found the

inclusion of T-intersections on graph elements to be useful, while others preferred them to be left out.

Discussion

The hypothesis that low complexity questions would be answered more accurately than high complexity questions was supported in the case of bar graph questions, but not for boxplot questions. One potential explanation for this difference is the sample's relative lack of experience with boxplots as compared to bar graphs, which relates Pinker's (1990) concept of *instantiated graph schema*, or knowledge of specific graph types – it would be predicted that increased experience with a particular type of graph would be associated with improved performance using that type of graph. However, participants who had created a boxplot or taken a statistics class did not outperform participants with less boxplot experience or those who had not taken a statistics class. Because boxplot experience questions were binary in nature, it is possible that they failed to capture the degree of boxplot experience, and that even participants who reported having used or created a boxplot had much less experience with this type of graph than with bar graphs.

The hypothesis that participants would perform similarly for low complexity tasks, regardless of data-ink level, was supported – for both bar graph and boxplot questions, no statistically significant differences in accuracy were found between data-ink conditions. The hypothesis that accuracy would increase with data-ink level in high complexity questions was not supported – for both bar graph and boxplots, no statistically significant differences in accuracy were found between data-ink conditions.

The hypothesis for increased accuracy in high complexity/high data-ink conditions was not supported. Accuracy in high data-ink/high complexity conditions was slightly lower than in low data-ink/high complexity and medium data-ink/high complexity conditions for both bar graph and boxplot questions, though those differences were not statistically significant. Contrary to this hypothesis, participants in the high data-ink/low complexity/bar graph condition performed significantly *more* accurately than those in the

high data-ink/high complexity/bar graph condition. In all other cases, data-ink and complexity did not have interactive effects on accuracy.

The hypothesis that high complexity questions would yield higher mental effort ratings than low complexity questions was supported. For both bar graph and boxplot questions, mean mental effort ratings were significantly higher in high complexity conditions than in low complexity conditions. Although these differences were small and there was only one statistically significant difference in accuracy between high and low complexity questions, this suggests that participants did in fact find the high complexity questions to be more difficult than low complexity questions, and provides further evidence for Carpendale's (2008) task complexity definitions.

The hypothesis that low complexity tasks would result in similar levels of mental effort, regardless of data-ink level was partially supported. For bar graph questions, mental effort ratings in the medium data-ink/low complexity condition were less variable than in the low and high data-ink conditions. The largest variability in mental effort ratings for low complexity bar graph and boxplot questions occurred in the high data-ink conditions. However, these differences were not large, and there were no statistically significant differences in mean mental effort ratings.

The hypothesis that increasing data-ink levels would result in decreasing levels of mental effort for high complexity questions was not supported. The pattern of mental effort responses for high complexity questions was generally similar to that of low complexity questions, with the widest range of mental effort ratings occurring in high data-ink conditions and the smallest range occurring in medium data-ink conditions. Again, these differences were relatively small and there were no statistically significant differences in mean mental effort ratings.

Although there is some support for data-ink maximization in the literature, those studies used small samples and misrepresented Tufte's recommendations. For example, Gillan and Richman (2009) claimed to have found limited support for data-ink

maximization, as participants used their low data-ink graphs less accurately and slower than medium and high data-ink graphs. However, only their medium and high data-ink graphs accurately reflected Tufte's recommendations, and like the present study, no performance differences were found between those designs. Although it is possible that Holmes-style graphs, which use visual metaphors, lead to performance deficits when participants are asked to make precise judgments about graphed data, that finding does *not* suggest that Tufte's recommendations to remove features like gridlines and T-intersections and to reduce the size of other graph elements are sound. Indeed, the results of this study suggest that those recommendations do not lead to improved performance.

Lohse's (1997) model of graph comprehension predicts that graph designs which require less effort to comprehend are those that shift processing from working memory to perceptual systems. The results of this study suggest that maximizing data-ink does *not* facilitate this shift. At best, mental effort ratings were similar, regardless of data-ink level, and the most variable mental effort ratings occurred in high data-ink conditions. At the same time, the presence of "chartjunk" in the form of gridlines and T-intersections did not seem to facilitate the comprehension processes described by Simkin and Hastie (1987), such as *projection*, or "sending out a ray" from one graph point to another, though some participants did presumably "miss" those features, and drew them on their graphs themselves.

Increases in the variability of mental effort responses for high data-ink graphs seem to make sense in light of interview data – multiple interviewees felt that the high data-ink graphs would be more difficult to interpret than the low or medium data-ink designs. However, the high data-ink graphs appeared to be as usable as other designs, and although high data-ink graphs received more variable mental effort ratings than low and medium data-ink graphs, participants were still able to answer questions with high data-ink graphs with similar levels of accuracy.

Both interview and experiment data suggest that if there *is* an optimal design, it may

be a medium data-ink level, as most interviewees preferred those designs and they yielded less variable mental effort ratings in many cases. It is also possible that this reduced mental effort variability was related to the graph lesson that occurred prior to participation, which used medium data-ink graphs when explaining bar graph and boxplot features. In either case, the effect was not large, and participants did *not* answer questions more accurately using medium data-ink graphs.

With regard to Tufte's claim that his high data-ink designs would be accepted with time, interview feedback indicated that high data-ink designs are not encountered or accepted by frequent users of graphs. Although models of graph comprehension and the results of the present study *do* seem to support the claim that viewers would be accustomed to high data-ink ratio designs, it does not seem that they have started to "catch on" in the years since Tufte published the data-ink concept.

Future Directions. As noted previously, *instantiated graph schemas* – knowledge regarding specific graph types – have been identified as an important factor in graph comprehension (Pinker, 1990). One interviewee commented that she would not have been able to identify Tufte's high data-ink boxplot as a boxplot without the context provided by the interview, suggesting that this particular graph did not activate a boxplot schema for that individual. Just as interviewees were informed of the type of graph they were being shown, experiment participants were told which types of graphs they would be using and given a lesson on interpreting both graph types prior to participation. Future studies should examine the effect of data-ink levels on graph comprehension with naive participants of varying experience levels. It is possible that some participants would have had more difficulty with the high data-ink boxplot had they received less pre-participation information about graph types.

Because none of the sample reported majoring in psychology or human factors, it seems unlikely that a large portion of participants was familiar with the type of studies from which the graphs came. However, this study did not attempt to measure graph

content knowledge, an important factor in models of graph comprehension. Future studies should investigate the impact of content knowledge as it relates to the data-ink ratio concept.

Multiple interviewees noted that they found the high data-ink bar graph to be too wide, a criticism which suggested that this graph did not make effective use of gestalt grouping principles. Future studies could make proportional reductions in the overall size of graphs as graph elements are reduced in size. For example, rather than holding the size of the graph constant, the proportion of white space to bar width could be held constant – as the size of bars is reduced, the overall width of the graph could be reduced. Similarly, one interviewee noted that the thin bars of the high data-ink bar graph made it more difficult to differentiate between bar colors, another criticism which suggested that the high data-ink bar graph did not fully leverage gestalt grouping principles. This criticism was similar in nature to the criticisms of Few (2009). However, neither of these criticisms were validated in the experiment – neither accuracy nor mental effort ratings suffered when using the high data-ink bar graph. Although it is possible that graph designs which balance data-ink maximization with more effective use of gestalt grouping principles could yield different results, the present findings suggest that participants are resilient to superficial design changes and would perform similarly regardless.

The present study used accuracy and mental effort as performance measures. It was assumed that increases in the amount of time participants required to understand graphs and answer questions would be reflected in reflected in mental effort ratings. Future studies could test this assumption by measuring or limiting response time and measuring response accuracy. It is possible that specific graph designs result in longer response times, and that these performance differences are not accurately captured by self-report data.

Although previous work has shown that design characteristics *do* affect comprehension, the design changes in this study did not. This suggest that not all display characteristics are created equally. Future studies should attempt to identify the design

characteristics which *do* affect performance in graph comprehension tasks. Additionally, future studies should further investigate the role of aesthetics and aesthetic preferences with regard to performance in graph comprehension tasks. Previous studies have measured preferences with regard to the data-ink ratio, but none have examined the relationship between performance and graph design preferences. It is possible that there is a connection between data-ink preferences and real or perceived benefits to performance in graph comprehension tasks.

Conclusion

The results of this study do not support claims that high data-ink ratio graph designs will result in increased graph comprehension. Rather, they suggest that the data-ink ratio concept deals with the subjective issue of graph aesthetics. Arguments about the aesthetics of graphs are worth having – interview data showed that graph creators care about the look of graphs and make efforts to ensure that their graphs meet their aesthetic standards. A graph creator who prefers the look of Tufte's high data-ink graphs should feel free to use them, but graph creators should *not* feel that maximizing data-ink ratio will result in more usable graphs. In defending his ideas, Tufte argued that it would be a mistake to underestimate the audiences of graphical information. With regard to graph designs with different data-ink ratios, this sentiment seems to be appropriate – graph users with varying levels of experience can extract complex information from high data-ink ratio designs. But they are just as good when data-ink ratio is not maximized. Future studies should further investigate the relationship between aesthetic preferences and the data-ink ratio.

References

Bateman, S., Mandryk, R., Gutwin, C., Genest, A., McDine, D., & Brooks, C. (2010). Useful junk? the effects of visual embellishment on compreshension and memorability of charts. *Proceedings of the SIGCHI Conference on Human Factors in Computing systems*, 2573–2582.

Beyer, H., & Holtzblatt, K. (1999, January). Contextual design. *Interactions*, *6*(1), 32–42.

Blasio, A. J., & Bisantz, A. M. (2002). A comparison of the effects of data-ink ratio on performance with dynamic displays in a monitoring task. *International Jounral of Industrial Ergonomics*, *30*, 89-101.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77-101.

Carpendale, S. (2008). Evaluating information visualizations. In A. Kerren, J. T. Stasko, J.-D. Fekete, & C. North (Eds.), *Information visualization* (p. 19-45). Berlin, Heidelberg: Springer-Verlag.

Carpenter, P. A., & Shah, P. (1998). A model of the perceptual and conceptual processes in graph comprehension. *Journal of Experimental Psychology: Applied*, *4*, 75–100.

Carswell, C. M. (1992). Choosing specifiers: An evaluation of the basic tasks model of graphical perception. *Human Factors*, *34*, 535-554.

Cleveland, W., & McGill, R. (1985). Graphical perception and graphical methods for analyzing scientific data. *Science*, *229*, 828–833.

Few, S. (2009). Sometimes we must raise our voices. *Visual Business Intelligence Newsletter*.

Freedman, E., & Shah, P. (2002). Toward a model of knowledge-based graph comprehension. In M. Hegarty, B. Meyer, & N. Narayanan (Eds.), *Diagrammatic representation and inference* (Vol. 2317, p. 18-30). Springer Berlin Heidelberg.

Fry, B. (2008). *Visualizing data*. Beijing: OReilly Media, Inc.

Gillan, D. J., & Richman, E. H. (1994). Miinimalism and the syntax of graphs. *Human*

Factors, *36*, 619-644.

Gillan, D. J., & Sorensen, D. (2009). Minimalism and the syntax of graphs ii: Effects of graph backgrounds on visual search. *Proceedings of the human factors and ergonomics society annual meeting*, *53*, 1096-1100.

Huang, W., Eades, P., & Hong, S.-H. (2009, June). Measuring effectiveness of graph visualizations: A cognitive load perspective. *Information Visualization*, *8*(3), 139–152.

Hullman, J., Adar, E., & Shah, P. (2011). Benifetting infovis with visual difficulties. *IEEE Transaction on Visualization and Computer Graphics*, *17*(12), 2213-2222.

Inbar, O., Tractinsky, N., & Meyer, J. (2007). Minimalism in information visualization: attitudes towards maximizing the data-ink ratio. *Proceedings of the 14th European conference on cognitive ergonomics: invent! explore!*, 185–188.

Katz, J. (2012). *Designing information: Human factors and common sense in information design.* Hoboken, NJ: Wiley.

Kelly, J. D. (1989). The data-ink ratio and accuracy of newspaper graphs. *Jounralism Quarterly*, *66*, 632–639.

Kirk, A. (2012). *Data visualization: a successful design process ; a structured design approach to equip you with the knowledge of how to successfully accomplish any data visualization challenge efficiently and effectively.* Birmingham, UK: Packt Pub.

Kosslyn, S. M. (1985). Graphics and human information processing: A review of five books. *Journal of the American Statistical Association*, *80*, 499-512.

Kosslyn, S. M. (2006). *Graph design for the eye and mind.* New York: Oxford University Press. Retrieved from `www.summon.com`

Kulla-Mader, J. (2007). *Graphs via ink: Understanding how the amount of non-data ink in a graph affects perception and learning.* (Unpublished master's thesis). University of North Carolina at Chapel Hill.

Lam, H., Bertini, E., Isenberg, P., Plaisant, C., & Carpendale, S. (2012). Empirical studies

in information visualization: Seven scenarios. *IEEE Transactions on Visualization and Computer Graphics*, *18*(9), 1520-1536. Retrieved from `www.summon.com`

Lellis, V. R. R., Mariani, M. M. d. C., Ribeiro, A. d. F., Cantiere, C. N., Teixeira, M. C. T. V., & Carreiro, L. R. R. (2013). Voluntary and automatic orienting of attention during childhood development. *Psychology and Neuroscience*, *6*(1), 15-21.

Levitin, D. J. (2011). *Foundation of cognitive psychology: Core readings*. Boston: Pearson Education.

Lohse, G. (1997). The role of working memory on graphical information processing. *Behaviour and Information Technology*, *16*, 297–308.

Paas, F. G. W. C. (1992). Traning strategies for attaining transfer of problem-solving skill in statistics: A cognitive load approach. *Journal of Educational Psychology*, *84*, 429-434.

Pinker, S. (1990). A theory of graph comprehension. In R. Friedle (Ed.), *Articificial intelligence and the future of testing* (p. 73-126). Hillsdale, NJ: Erlbaum.

Portigal, S. (2013). *Interviewing users: how to uncover compelling insights*. Brooklyn, New York: Rosenfeld Media. Retrieved from `www.summon.com`

Rit in brief. (2015, April 10). Retrieved from `https://www.rit.edu/overview/rit-in-brief`

Robertson, G., Czerwinski, M., Fisher, D., & Lee, B. (2009). Chapter 2 selected human factors issues in information visualization. *Reviews of Human Factors and Ergonomics*, *5*(1), 41-41.

Romoser, M. R. E., & Fisher, D. L. (2009). The effect of active versus passive training strategies on improving older drivers' scanning in intersections. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *51*(5), 652-652.

Shah, P. (1995). Cognitive processes in graph comprehension [Doctoral Dissertation]. *ProQuest Dissertations and Theses*, 1-162.

Shah, P., Freedman, E. G., & Vekiri, I. (2005). The comprehension of quantitative

information in graphical displays. In P. Shah & A. Miyake (Eds.), *Cambridge handbook of visuospatial thinking.* New York: Cambridge University Press.

Shah, P., Mayer, R. E., & Hegarty, M. (1999). Graphs as aids to knowledge construction: Signaling techniques for guiding the process of graph comprehension. *Journal of Educational Psychology*, *91*(4), 690 - 702.

Simkin, D., & Hastie, R. (1987). An information processing analysis of graph perception. *Journal of the American Statistical Association*, *82*, 454–465.

Stewart, B. M., Cipolla, J. M., & Best, L. A. (2009). Extraneous information and graph comprehension. *Campus-Wide Information Systems*, *26*(3), 191-200.

Tory, M., & Moller, T. (2004). Human factors in visualization research. *IEEE Transactions on Visualization and Computer Graphics*, *10*(1), 72-84.

Tory, M., & Moller, T. (2005). Evaluating visualizations: Do expert reviews work? *IEEE Computer Graphics and Applications [H.W. Wilson - AST]*, *25*(5), 8-11.

Tufte, E. (1983). *The visual display of quantitative information.* Cheshire, CT: Graphics Press.

Tufte, E. (2001). *The visual display of quantitative information.* Cheshire, CT: Graphics Press.

Tufte, E. (2015, March 11). *Analytical design and human factors.* Retrieved from `www.edwardtufte.com/bboard/q-and-a-fetch-msg?msg_id=0000KI`

Tukey, J. W. (1990). Data-based graphics: Visual display in the decades to come. *Statistical Science*, *5*(3), 327–339.

Wainer, H. (1984). How to display data badly. *The American Statistician*, *38*(2), 136-147.

Wickens, C., & Holland, J. (2000). *Engineering psychology and human performance.* Upper Saddle River, NJ: Prentice Hall.

Zhu, Y. (2007). Measuring effective data visualization. In G. Bebis et al. (Eds.), *Advances in visual computing* (Vol. 4842, p. 652-661). Springer Berlin Heidelberg.

Appendix A

Paas Mental Effort Scale – Original and Adapted

Please rate the mental effort required to answer question #1

1 = very, very low mental effort
2 = very low mental effort
3 = low mental effort
4 = rather low mental effort
5 = neither low nor high mental effort
6 = rather high mental effort
7 = high mental effort
8 = very, very high mental effort
9 = extremely high mental effort

Figure A1. Original version of the Paas mental effort scale, which used different language at the extreme ends of the scale.

Please rate the mental effort required to answer question #1

1 = very, very low mental effort
2 = very low mental effort
3 = low mental effort
4 = rather low mental effort
5 = neither low nor high mental effort
6 = rather high mental effort
7 = high mental effort
8 = very high mental effort
9 = very, very high mental effort

Figure A2. Adapted version of the Paas mental effort scale, which appeared after every question. A rating of 8 was changed to "very high mental effort" and 9 was changed from "extremely high" to "very, very high mental effort" to match the low end of the scale.

Appendix B

Graph Explanation Paragraphs

Bar Graph

The following graph depicts the results of an experiment which measured reaction time in an attention orienting task – which asked children from 5 different age groups to respond with a key press after the appearance of a stimulus. The experimental stimulus – a solid black box – was presented at particular intervals. The box was considered either valid – in a location indicated by an arrow-shaped cue – or invalid – opposite the location indicated by the arrow-shaped cue. Participants were asked to respond to the black box as soon as possible by pressing the spacebar, regardless of whether it was valid or invalid. Reaction times were measured in milliseconds.

Boxplot

The following graph depicts the results of an experiment which measured secondary looks – the act of glancing in the most likely direction of oncoming traffic after beginning a turn – in a group of older drivers. All participants completed pre-training simulator and field tests in which the proportion of intersections with secondary looks was measured. Next, one group received active, immersive training using the simulator while a second group received passive, classroom-style training. A control group received no further training. All 3 groups were then tested again in both the simulator and in the field, with the proportion of intersections with secondary looks measured again.

Appendix C

Graph Lesson

The italicized text accompanying each image was presented verbally to participants as the image displayed.

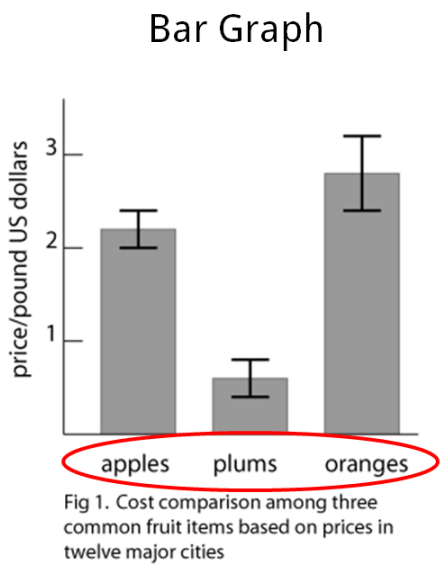


Fig 1. Cost comparison among three common fruit items based on prices in twelve major cities

- *Bar graphs typically show comparisons between categories. The categories are arranged along the X-axis.*
- *In this graph, the categories are fruit types – apples, plums, oranges.*
- *The measure by which categories are being compared is on the Y-axis. On this graph the measure is price per pound in US dollars.*

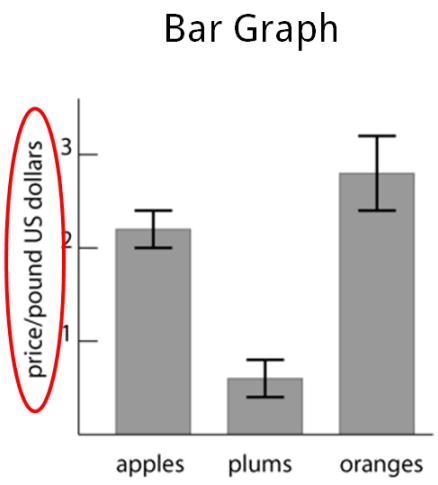


Fig 1. Cost comparison among three common fruit items based on prices in twelve major cities

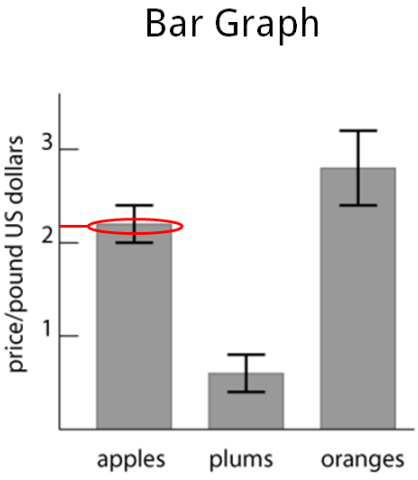


Fig 1. Cost comparison among three common fruit items based on prices in twelve major cities

- *The bars themselves have lengths proportional to the values they represent, which are typically means.*
- *The top edge of the bar represents the value for that category. The average price per pound of apples in this graph is around $2.10.*
- *The darker lines at the end of each bar are error bars. They typically reprsent different measures of variability.*

Bar Graph

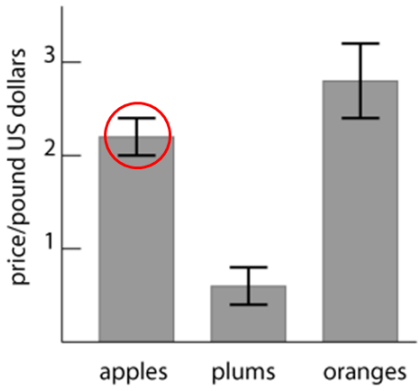
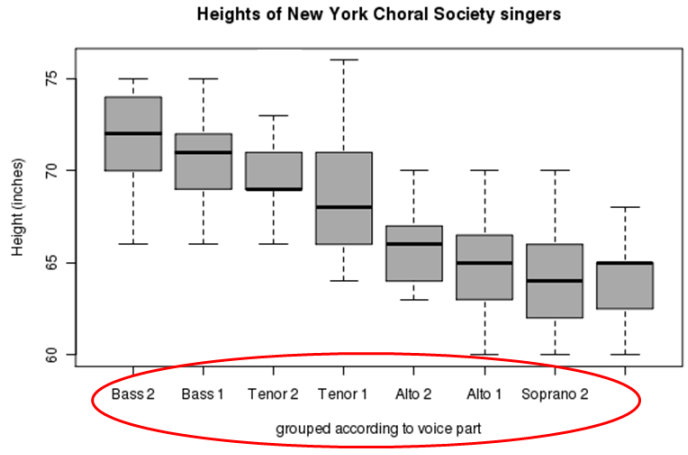


Fig 1. Cost comparison among three common fruit items based on prices in twelve major cities

- *In the bar graph you'll be looking at, they will represent the standard error of the mean. If the error bars of two categories overlap, it's unlikely that there is a statistically significant different between them.*

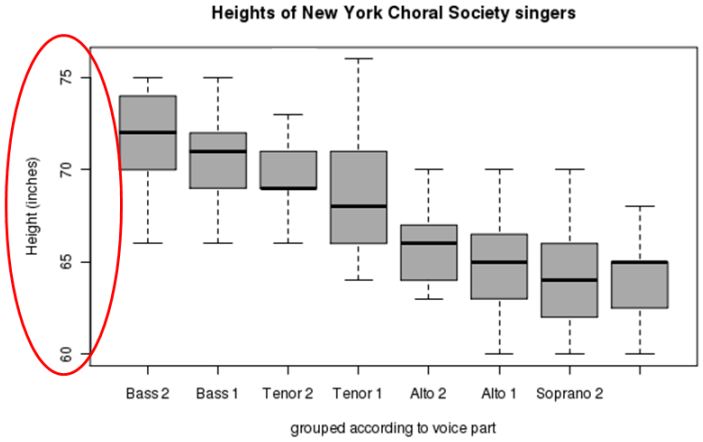
Box Plot



- *Boxplots also show comparisons between categories. The categories are arranged along the X-Axis.*

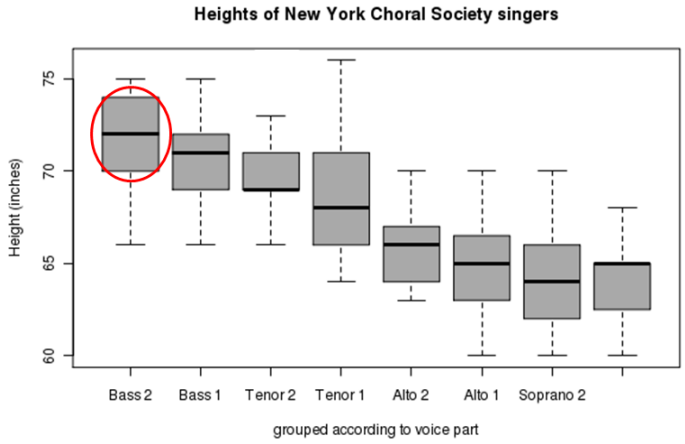
- *In this graph, the categories are voice parts in a chorus âĂŞ bass, tenor, etc.*

Box Plot



- *Again, the measure by which the categories are being compared is on the Y-Axis. On this graph, the measure is height in inches.*

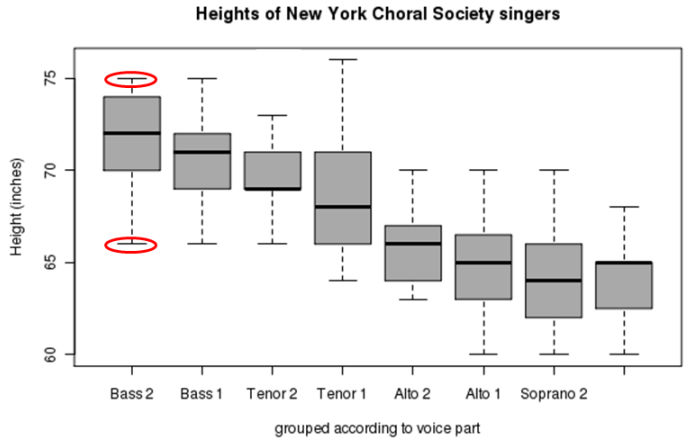
Box Plot



- *Boxplots represent data through quartiles, which divide a data set into 4 groups, each representing 25% of the data. The box represents the middle 50% of the data set.*
- *The upper edge of the box represents the value at which 25% of the data is greater. 25% of bass 2 singers are taller than around 74 inches.*

- *The bottom edge of the box represents the value at which 25% of the data is smaller. 25% of bass 2 singers are shorter than around 70 inches.*
- *The dark line inside the box represents the median âĂŞ the middle value in the data set. The median bass 2 singer is around 72 inches tall.*

Box Plot



- *The lines protruding from the top and bottom of the box are called whiskers, and can represent different things. In the graph you will look at, they represent the maximum and minimum values in the dataset.*
- *The tallest singer in the bass 2 group is around 75 inches.*
- *The shortest singer in the bass 2 group is around 66 inches.*

Appendix D

Graph Comprehension Questions and Answers

Answers to questions presented in italics.

Low Complexity Questions – Bar Graph

1. Which variable had a larger effect on reaction time – age or validity?

   Age

2. Which age group had the fastest reaction times?

   Age 10

High Complexity Questions – Bar Graph

1. Based on the graph, what reaction times (in milliseconds) would you predict for a group of 12 year olds:

   Valid stimuli –

   Invalid stimuli –

   Correct answers predicted reaction times between 300-420 for valid stimuli and 300-450 for invalid stimuli. Predictions in which the invalid reaction time was lower than valid were considered incorrect.

2. What can you conclude about the effect of the valid/invalid condition on reaction time?

   Invalid condition leads to slower reaction times AND/OR the difference between reaction times goes down with age.

Low Complexity Questions – Boxplot

1. In the passive learning/field drive condition, did the higher maximum proportion of intersections with secondary looks occur before or after training?

   Before / Pre-Training.

2. Did training have an effect in the passive learning/field drive condition?

   No. Answers such as "very little," "barely," etc. were also accepted.

High complexity Questions – Boxplot

1. Which learning method (active vs. passive & simulator vs. field drive) would you recommend for increasing safe driving habits?

   Active learning/field drive. Answers that did not include both parts were not accepted.

2. The experimenters decide to assess a third training style that blends active and passive learning. What would you predict to be the minimum proportion of intersections with secondary looks measured pre-training?

   Answers $\leq$ 15 were accepted. If participants gave a range of numbers, an average number was calculated for analysis.

Appendix E

Demographic Questionnaire

Are you:

Female **Male**

Age: _____

Year of study:

1st Year **2nd Year** **3rd Year** **4th Year** **5th Year or higher**

What is your major?: _____

No major

Have you ever taken a statistics class?

Yes **No**

Please respond to the following statements:

Before today, I had seen a bar graph.	**Yes**	**No**
Before today, I had used a bar graph.	**Yes**	**No**
Before today, I had made a bar graph.	**Yes**	**No**
Before today, I had seen a boxplot.	**Yes**	**No**
Before today, I had used a boxplot.	**Yes**	**No**
Before today, I had made a boxplot.	**Yes**	**No**

Figure E1. Demographic questionnaire included at the end of the experimental packet.

Appendix F

Pre-Interview Questions

1. How frequently do you use graphs? (in classes? In research?)
2. For what reasons do you use graphs?
3. What type(s) of graphs do you use most frequently? (Bar graphs? Scatterplots? etc.)
4. Are there types of graphs that you prefer?
5. How frequently do you create your own graphs?
6. What types of graphs do you create most frequently?
7. How do you create your own graphs? (What software do you use?)
8. Are there any software/tools you prefer to use?
9. What are the most important factors in the design of graphs that you create?

Appendix G

Example Discussion Guide

Interview #1

- Consent information
- Thesis deals with graphs and how graph design impacts graph interpretation.
- 2 parts:
  - Experiment: large sample of undergraduates, different designs, answering questions
  - Interviews: with experts (RIT faculty) from psychology, design, stats, etc.
- Goal of speaking with you is to understand how someone w/ expertise with graphs and data understands, uses and creates graphs.
- First, talk about your responses to pre-interview (graph use, graph creation)
- Also talk about a specific design guideline called the Data-Ink Ratio, ask for your feedback about it, and talk about some example graphs which illustrate the DIR concept.
- Should take around 1 hour
- Recording permission

For what reasons do you use graphs?
To learn or communicate relationships between different things (variables, system components, &c.)

- Look at graphs early on when reading an article?
- When you first look at a graph, what features do you focus on?

How frequently do you use graphs? (in classes? In research?)
Daily, I try to show different graphics in every lecture and put graphics in every handout I write. In research I always want to plot my data in many ways to literally see what is going on. I use graphics others have made to gain information about relationships between things, and I use graphics that I create myself to communicate information to others.

- Beyond showing graphs in courses, do you include content about graph *design* in your courses?

Are there types of graphs that you prefer?
I am a big fan of box plots because they contain such wealth of information about the distributions of data. In general, I prefer plots that show every data point; thus, bar graphs that only show the mean as the bar height are my least favorite

How frequently do you create your own graphs?
Almost daily. If I am in a middle of data analysis, I make dozens of plots in one sitting. Ditto for writing handouts or making lecture slides. I doodle with graphs just for fun, too, to wrap my head around things I am studying.

- Tell me more about graph doodling?
- In terms of design, are the graphs you create for data analysis/exploration different than graphs you would create for publication?

What types of graphs do you create most frequently?
Box plots, line graphs with 95% confidence intervals, scatterplots, histograms, concept maps, flowcharts.

How do you create your own graphs? (What software do you use?)
R for statistical graphs, CmapTools for concept maps and flowcharts. I recently learned of a new, free, software called yEd.

Are there any software/tools you prefer to use?
R is immensely powerful and allows for true creativity in presenting data in almost any imaginable way, but it has a shallow learning curve and any level of mastery requires much practice. CmapTools is simple and easy to use, but it has some limitations. I have yet to use yEd for any serious work.

- R is powerful in that it allows for lots of types of graphs? Freedom in design of graphs? Both?
- What is your experience with other graph creation tools?
- What is it about creating graphs in R that makes it "worth" the extra effort it requires?
- Do you use design templates to create graphs?
- Are there graph features (like gridlines for example) that you would use in some situations but not in others?

What are the most important factors in the design of graphs that you create?
Accuracy: The plot should reveal the relationships as they truly are and not obscure any phenomena.

Ease of use: The gist of the information conveyed by the graph should be immediately and intuitively perceptible.

- What sort of features/design choices would make a graph "inaccurate"
- What sort of features/design choices would obscure phenomena?
- What sort of features/design choices make a graph more intuitive?
- Do aesthetics play a role in "ease of use"?

I also believe there exists a proper match between data and graphs, e.g., lines for trend information, side-by-side boxes to compare two conditions or groups; such "natural mappings" should be exploited in graph design.

- Is a mismatch between data type and graph type something you encounter in journals, etc.?

Data-Ink Ratio

2 types of data – *data-ink* (parts of the graph that represent the data itself) and *non-data-ink* (ink that does *not* depict statistical information – "*chartjunk*"). Ratio = data-ink/non-data-ink. Can range from 0 to 1.

Chartjunk should almost always be removed, and size/weight graph elements should be reduced. Results in minimalist, spare graph designs.

Argued that higher data-ink ratios are better (faster and more accurate judgments).

- Heard of this guideline?
- (If so, do you have an opinion on it?)

Example Graphs

- Boxplots/Bar Graphs – do any of the designs make the boxplot/bar graph more/less effective? (Take longer to understand or process?)
- Are there features of any of these graphs that you consider to be "chartjunk?"
- Would the differences in design between the graphs have any effect on the ***accuracy*** of judgments you could make using them?
- Do you feel you process any of the graph designs ***faster*** or slower than the others?
- Outside of materials about the data-ink ratio, have you encountered graph designs that looked like High DIR graphs?
- Which design style is most similar to the graphs you encounter?
- Do you think that, with time, you would "get used" to the high data-ink graphs?
- Thinking about the graphs you see/use most frequently (time series), do those graphs typically resemble one of the examples?
- Based on high DIR graphs, if you were to imagine the high DIR version of time series, would you like that graph?

Appendix H

Results Tables

Table H1

Accuracy and mental effort data for bar graphs questions broken down by data-ink and complexity.

	Accuracy			Mental Effort		
	M	*SD*	95% CI	*M*	*SD*	95% CI
All questions	0.79	0.22	[.75, .82]	3.10	1.12	[2.92, 3.26]
Low complexity bar Qs	0.84	0.27	[.80, .88]	2.58	1.22	[2.39, 2.76]
High complexity bar Qs	0.74	0.31	[.69, .78]	3.61	1.28	[3.42, 3.80]
Low Data-ink Ratio	0.80	0.23	[.74, .86]	3.10	1.17	[2.76, 3.38]
Medium Data-ink Ratio	0.78	0.23	[.72, .84]	3.00	0.83	[2.79, 3.23]
High Data-ink Ratio	0.78	0.20	[.73, .84]	3.20	1.33	[2.84, 3.55]
Low DIR x low complexity	0.82	0.28	[.75, .90]	2.64	1.26	[2.30, 2.98]
Medium DIR x low complexity	0.82	0.29	[.74, .89]	2.54	1.00	[2.28, 2.81]
High DIR x low complexity	0.87	0.24	[.81, .93]*	2.54	1.40	[2.17, 2.92]
Low DIR x high complexity	0.77	0.31	[.69, .86]	3.49	1.33	[3.14, 3.84]
Medium DIR x high complexity	0.74	0.30	[.66, .82]	3.48	1.01	[3.22, 3.75]
High DIR x high complexity	0.70	0.32	[.61, .78]*	3.85	1.45	[3.47, 4.23]

Table H2

Accuracy and mental effort data for boxplot questions broken down by data-ink and complexity.

	Accuracy			Mental Effort		
	M	*SD*	95% CI	*M*	*SD*	95% CI
All questions	0.55	0.32	[.50, .60]	3.29	1.19	[3.11, 3.47]
Low complexity boxplot Qs	0.53	0.40	[.47, .59]	3.01	1.32	[2.81, 3.21]
High complexity boxplot Qs	0.57	0.36	[.51, .62]	3.57	1.32	[3.38, 3.78]
Low Data-ink Ratio	0.57	0.31	[.49, .65]	3.40	1.12	[3.12, 3.72]
Medium Data-ink Ratio	0.58	0.31	[.50, .66]	3.10	1.06	[2.83, 3.37]
High Data-ink Ratio	0.50	0.35	[.41, .59]	3.36	1.37	[3.00, 3.73]
Low DIR x low complexity	0.55	0.41	[.44, .66]	3.17	1.30	[2.82, 3.51]
Medium DIR x low complexity	0.56	0.38	[.46, .66]	2.87	1.13	[2.58, 3.16]
High DIR x low complexity	0.47	0.41	[.37, .58]	3.00	1.52	[2.59, 3.41]
Low DIR x high complexity	0.58	0.33	[.49, .67]	3.68	1.17	[3.36, 3.99]
Medium DIR x high complexity	0.59	0.36	[.50, .69]	3.33	1.29	[3.00, 3.67]
High DIR x high complexity	0.53	0.40	[.42, .63]	3.73	1.47	[3.33, 4.12]

Table H3

Accuracy and mental effort data for individual graph comprehension questions. Complexity levels are noted parenthetically.

	Accuracy			Mental Effort		
	M	*SD*	95% CI	*M*	*SD*	95% CI
Bar graph Q1 (high)	0.78	0.42	[.71, .84]	3.88	1.52	[3.65, 4.11]
Bar graph Q2 (low)	0.85	0.36	[.79, .90]	3.23	1.62	[2.99, 3.48]
Bar graph Q3 (high)	0.70	0.46	[.63, .77]	3.34	1.62	[3.09, 3.58]
Bar graph Q4 (low)	0.83	0.14	[.77, .88]	1.92	1.25	[1.73, 2.11]
Boxplot Q1 (low)	0.51	0.50	[.43, .58]	3.36	1.66	[3.11, 3.61]
Boxplot Q2 (low)	0.55	0.50	[.48, .63]	2.66	1.31	[2.46, 2.85]
Boxplot Q3 (high)	0.55	0.50	[.48, .63]	2.76	1.45	[2.54, 2.97]
Boxplot Q4 (high)	0.58	0.50	[.51, .65]	4.39	1.79	[4.12, 4.66]

Table H4

Accuracy and mental effort data for individual bar graph comprehension questions, broken down by data-ink level. Complexity levels are noted parenthetically.

		Accuracy			Mental Effort		
		M	*SD*	95% CI	*M*	*SD*	95% CI
Bar graph Q1 (high)							
	Low DIR	0.82	0.38	[.72, .93]	3.68	1.49	[3.29, 4.08]
	Medium DIR	0.78	0.42	[.68, .89]	3.83	1.37	[3.47, 4.19]
	High DIR	0.72	0.45	[.61, .84]	4.12	1.68	[3.68, 4.57]
Bar graph Q2 (low)							
	Low DIR	0.82	0.38	[.72, .93]	3.30	1.59	[2.88, 3.72]
	Medium DIR	0.80	0.40	[.70, .90]	3.31	1.51	[2.91, 3.71]
	High DIR	0.91	0.28	[.84, .99]	3.09	1.76	[2.62, 3.55]
Bar graph Q3 (high)							
	Low DIR	0.72	0.45	[.60, .84]	3.30	1.66	[2.86, 3.74]
	Medium DIR	0.70	0.46	[.58, .82]	3.14	1.43	[2.76, 3.51]
	High DIR	0.67	0.47	[.55, .80]	3.58	1.76	[3.11, 4.05]
Bar graph Q4 (low)							
	Low DIR	0.82	0.38	[.72, .93]	1.98	1.32	[1.63, 2.34]
	Medium DIR	0.83	0.38	[.74, .93]	1.78	0.97	[1.52, 2.03]
	High DIR	0.83	0.38	[.73, .93]	2.00	1.41	[1.62, 2.38]

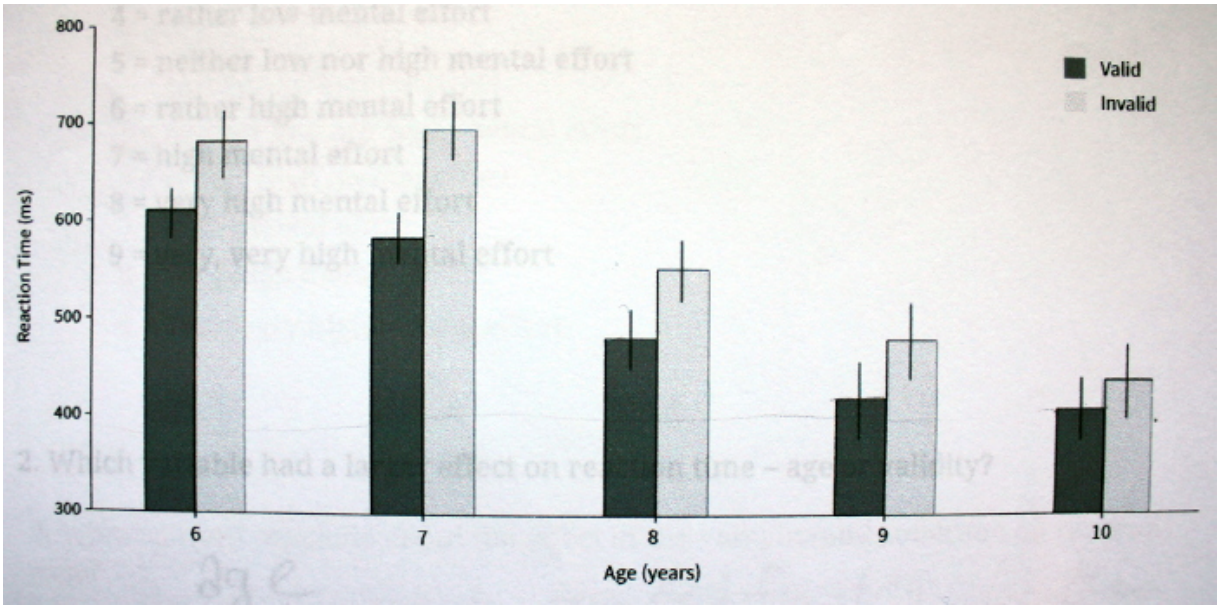
Table H5

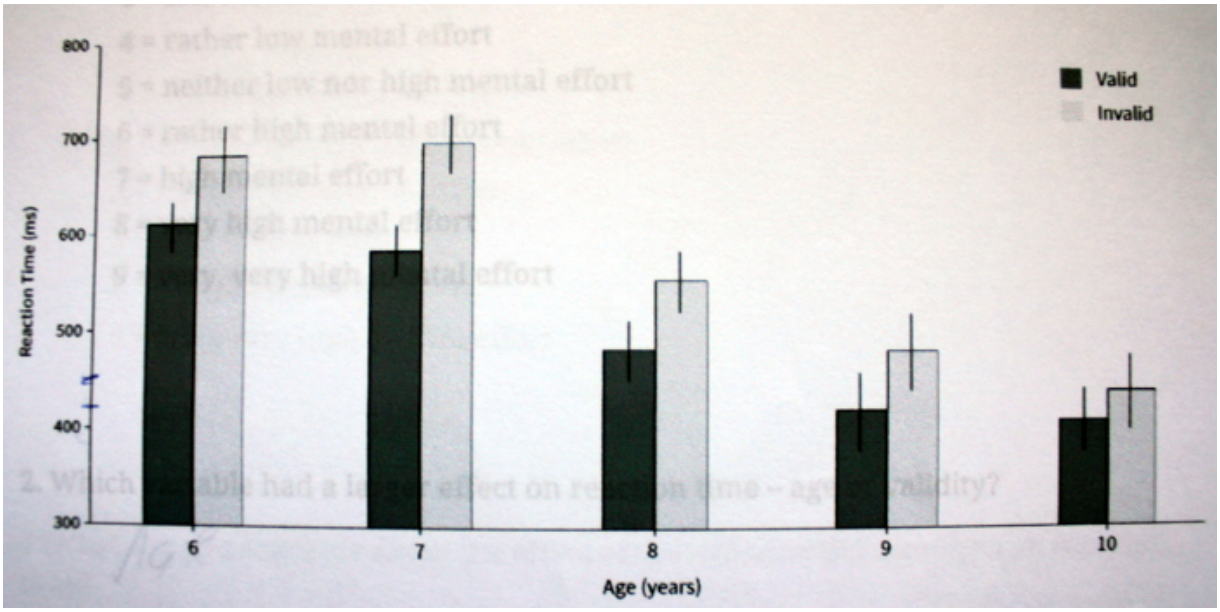
Accuracy and mental effort data for individual boxplot comprehension questions, broken down by data-ink level. Complexity levels are noted parenthetically.

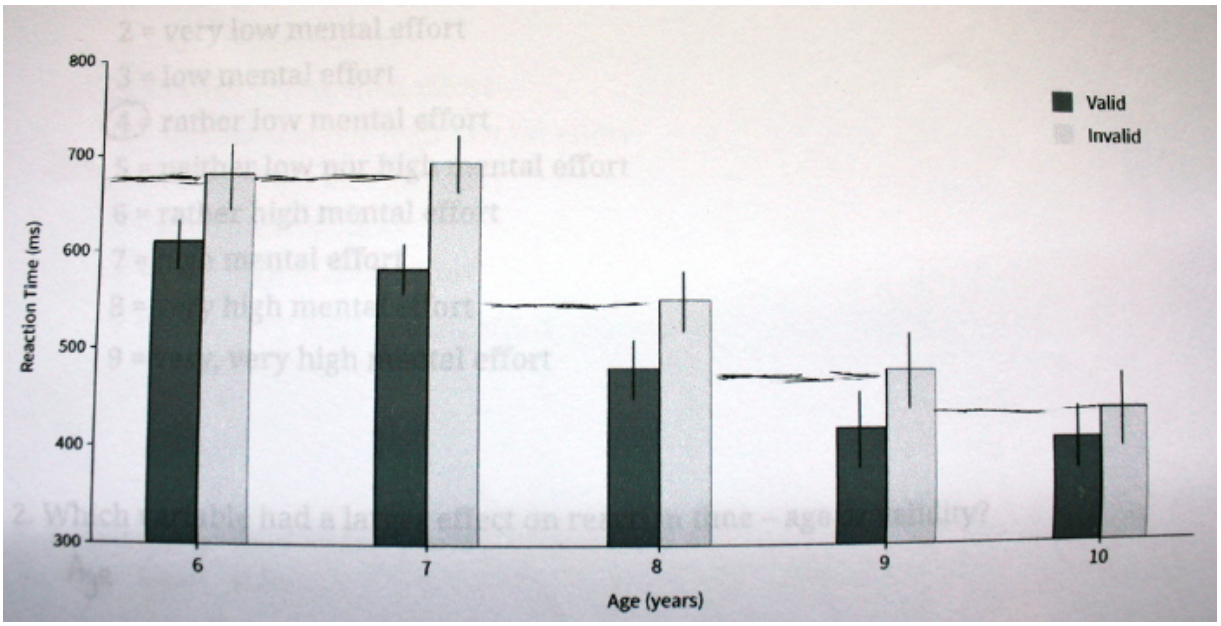
		Accuracy			Mental Effort		
		M	*SD*	95% CI	*M*	*SD*	95% CI
Boxplot Q1 (low)							
	Low DIR	0.52	0.50	[.38, .65]	3.58	1.69	[3.13, 4.03]
	Medium DIR	0.48	0.50	[.35, .61]	3.20	1.55	[2.80, 3.60]
	High DIR	0.52	0.50	[.38, .65]	3.30	1.74	[2.84, 3.77]
Boxplot Q2 (low)							
	Low DIR	0.59	0.50	[.46, .72]	2.75	1.24	[2.42, 3.08]
	Medium DIR	0.63	0.49	[.51, .76]	2.53	1.13	[2.24, 2.82]
	High DIR	0.43	0.50	[.30, .56]	2.70	1.54	[2.28, 3.11]
Boxplot Q3 (high)							
	Low DIR	0.55	0.50	[.42, .69]	2.89	1.19	[2.58, 3.21]
	Medium DIR	0.53	0.50	[.40, .66]	2.63	1.39	[2.27, 2.99]
	High DIR	0.57	0.50	[.44, .70]	2.75	1.73	[2.29, 3.21]
Boxplot Q4 (high)							
	Low DIR	0.61	0.49	[.48, .74]	4.46	1.76	[3.99, 4.92]
	Medium DIR	0.65	0.48	[.53, .77]	4.03	1.70	[3.59, 4.47]
	High DIR	0.48	0.50	[.35, .62]	4.70	1.89	[4.19, 5.20]

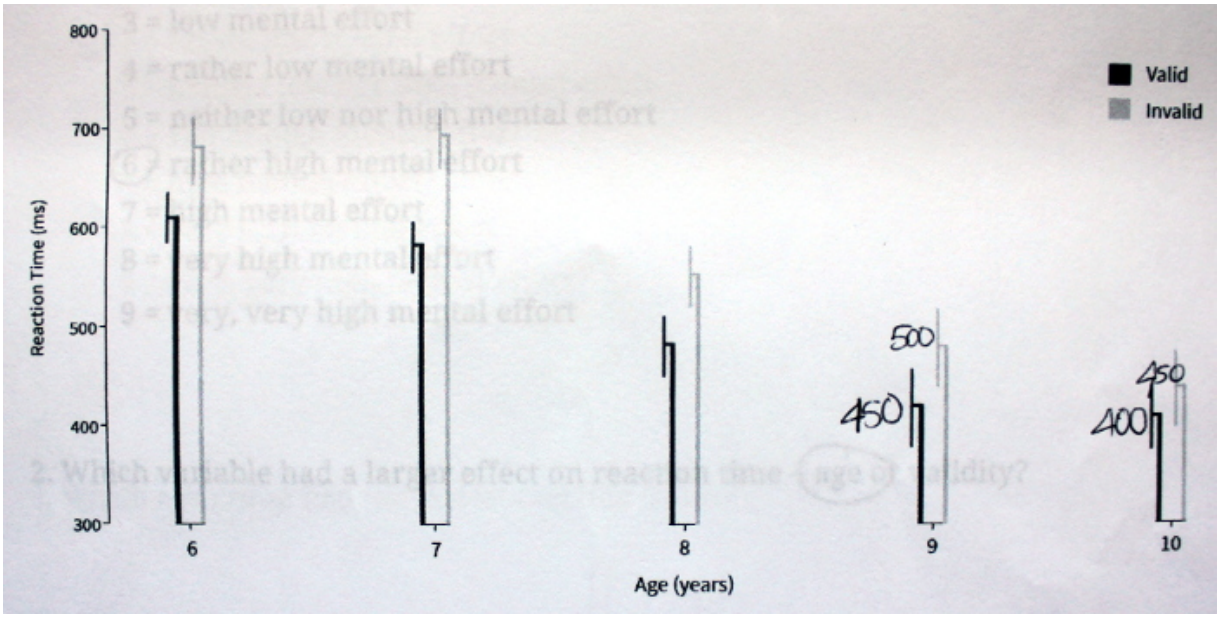
Appendix I

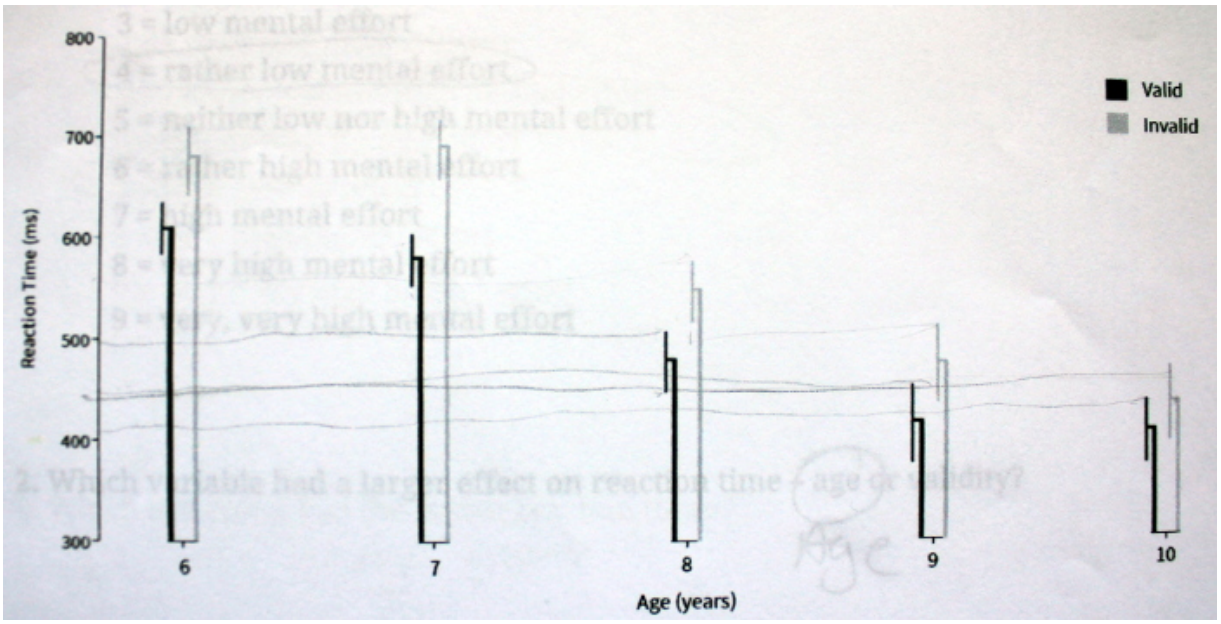
Graph Annotations

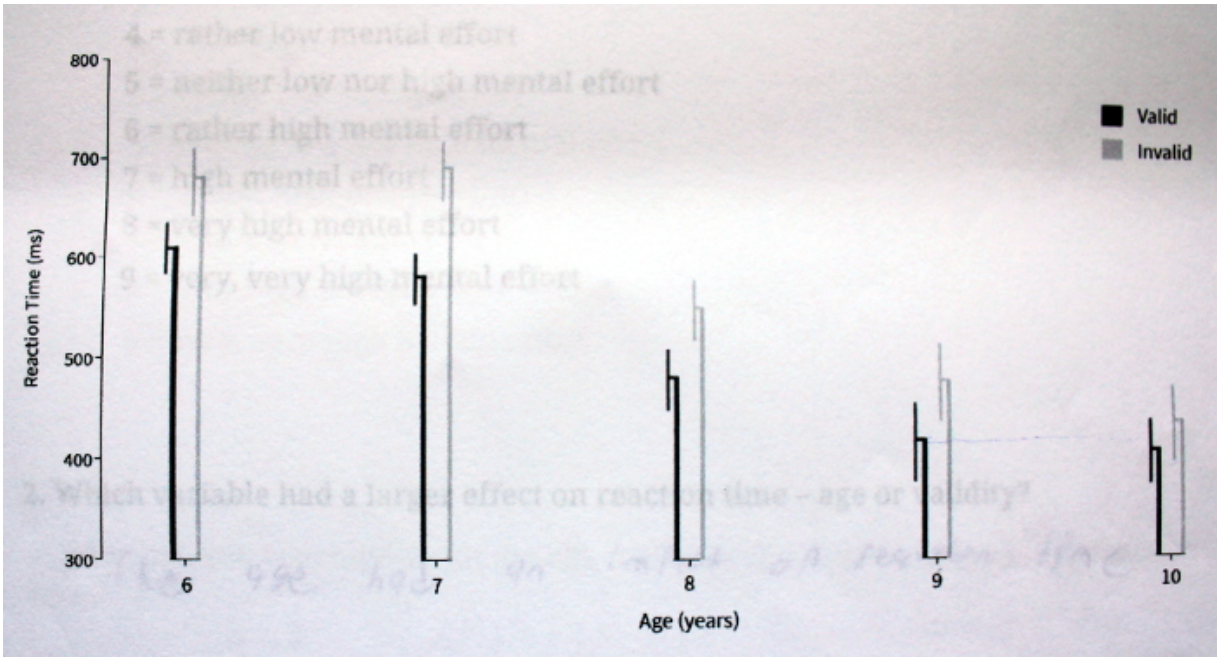


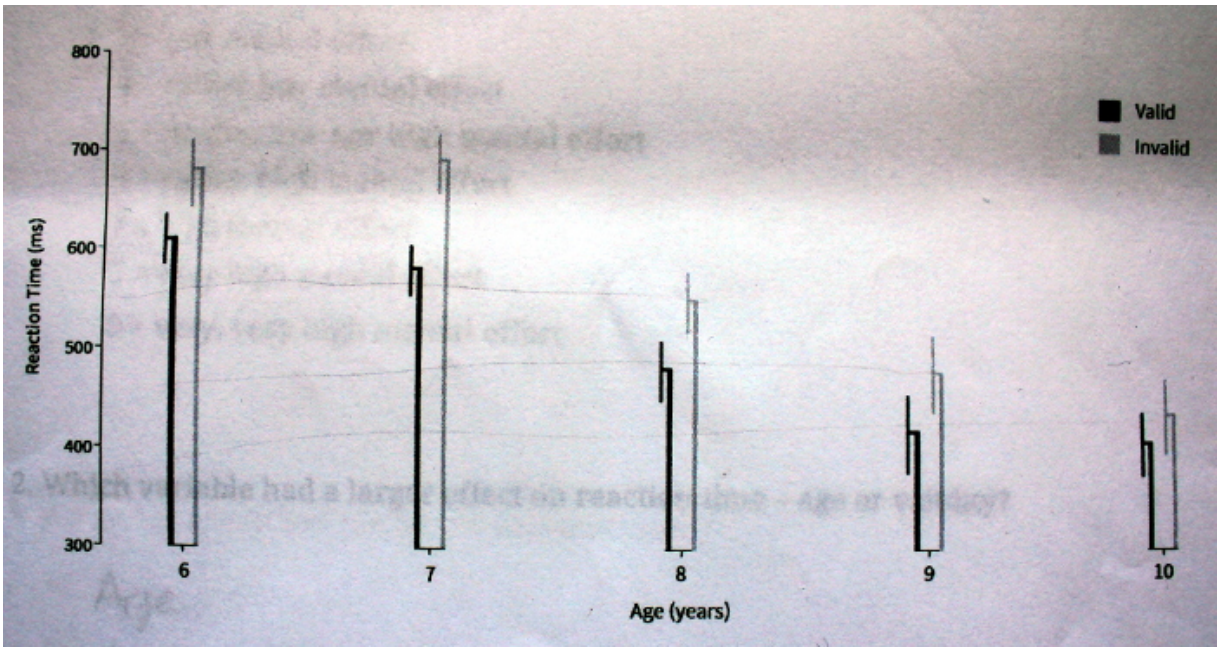


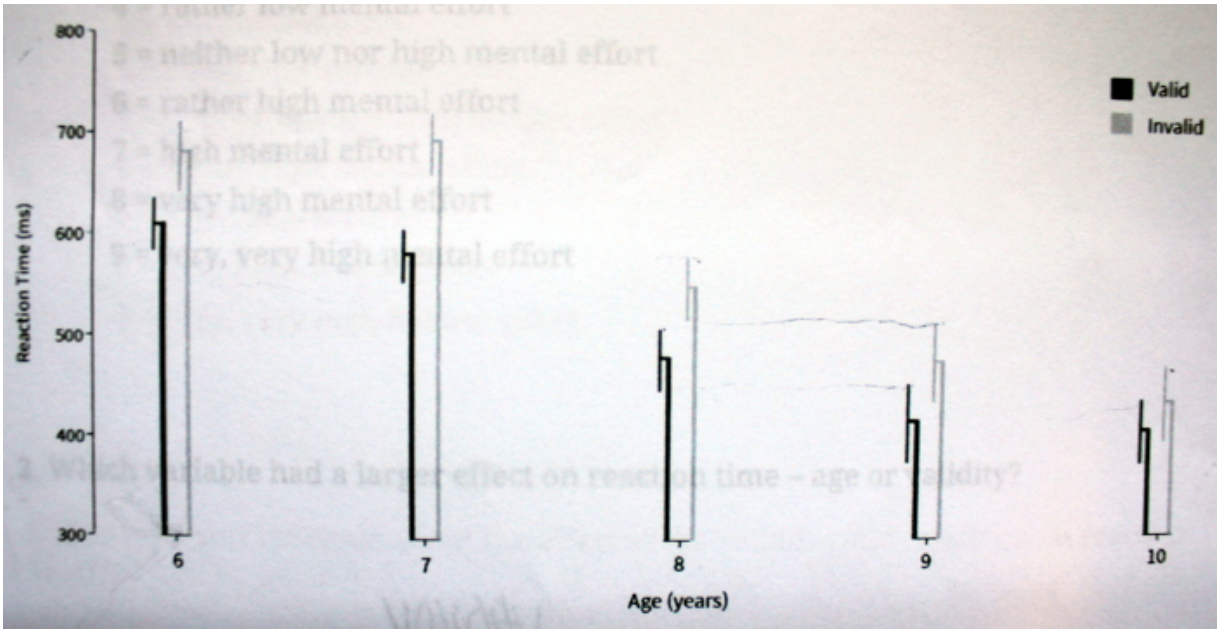

800
700
600
500
400
300
Reaction Time (ms)
Valid
Invalid
6
7
8
9
10
Age (years)

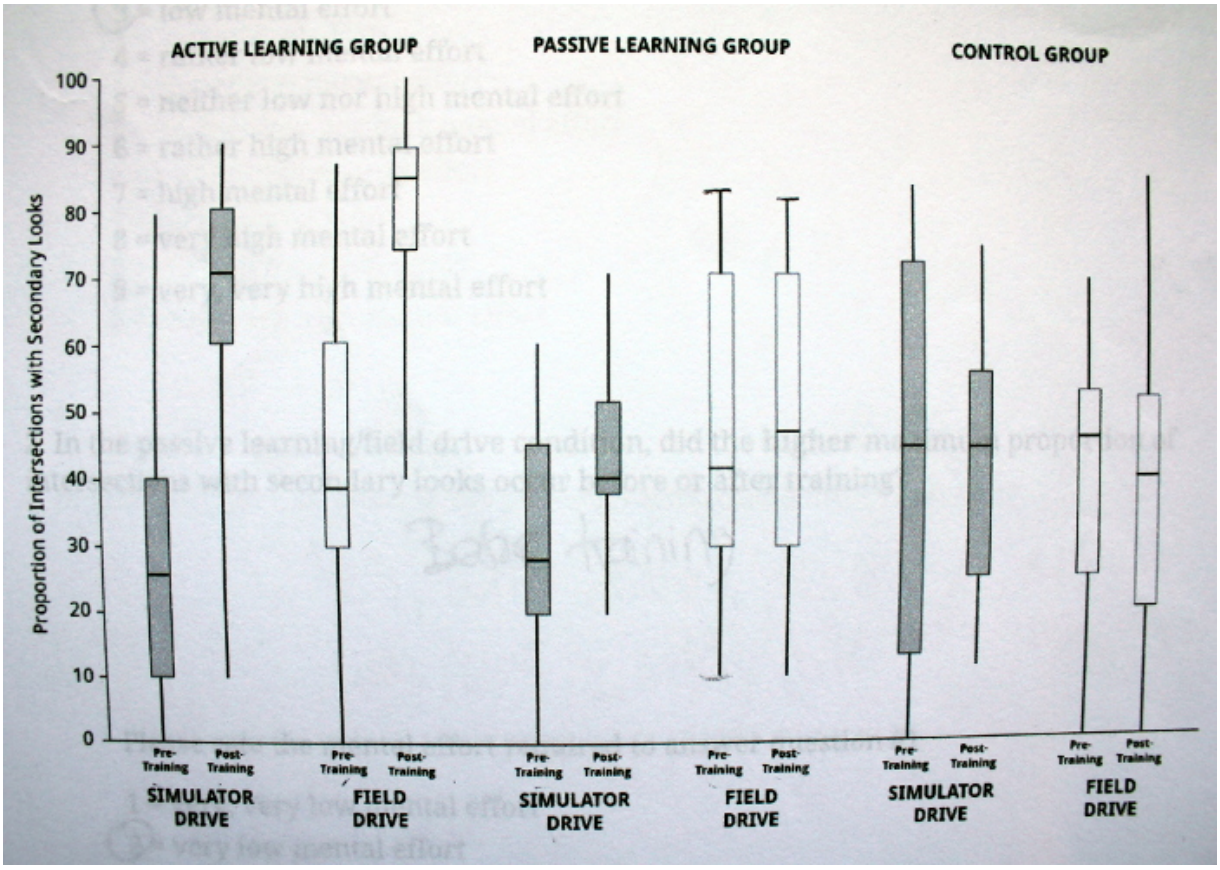

800
700
600
500
400
300
Reaction Time (ms)
Valid
Invalid
500
450
450
400
6
7
8
9
10
Age (years)

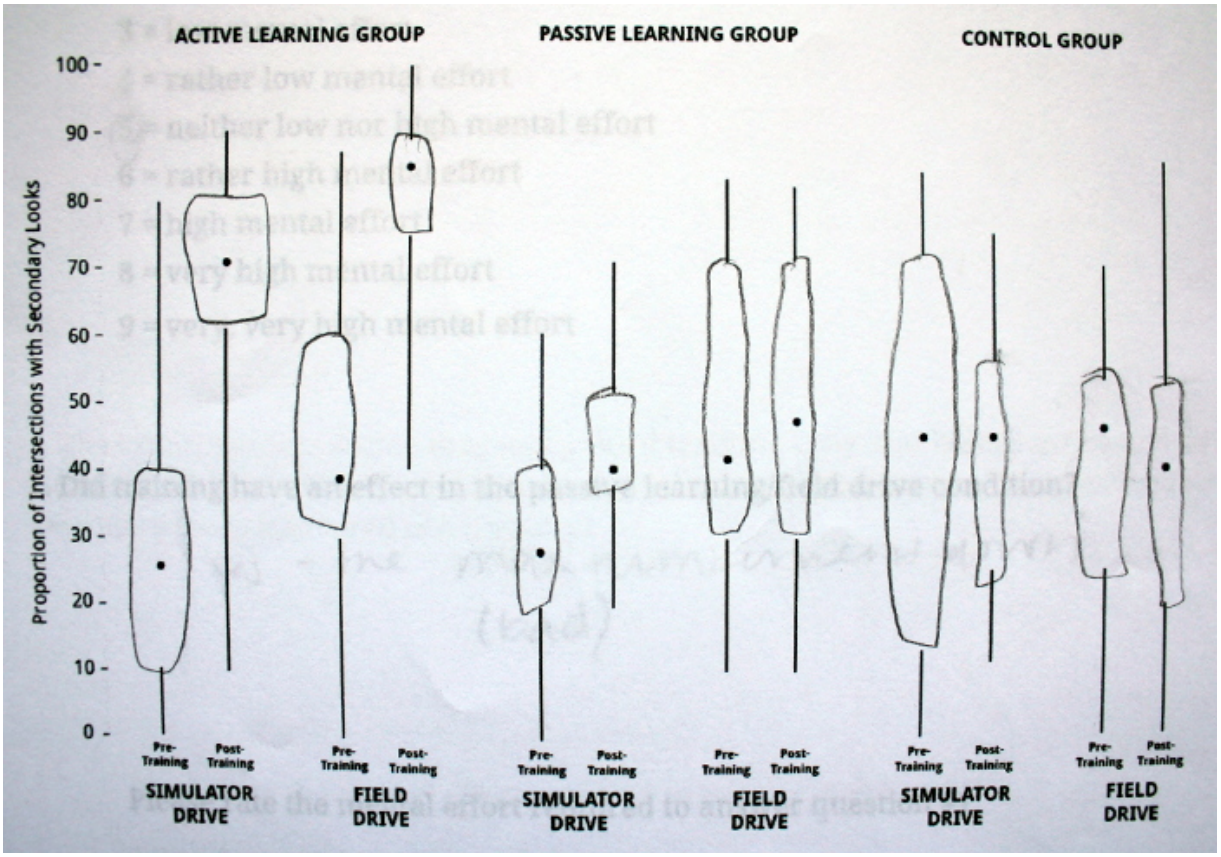

800
700
600
500
400
300
Reaction Time (ms)
Valid
Invalid
6
7
8
9
10
Age (years)

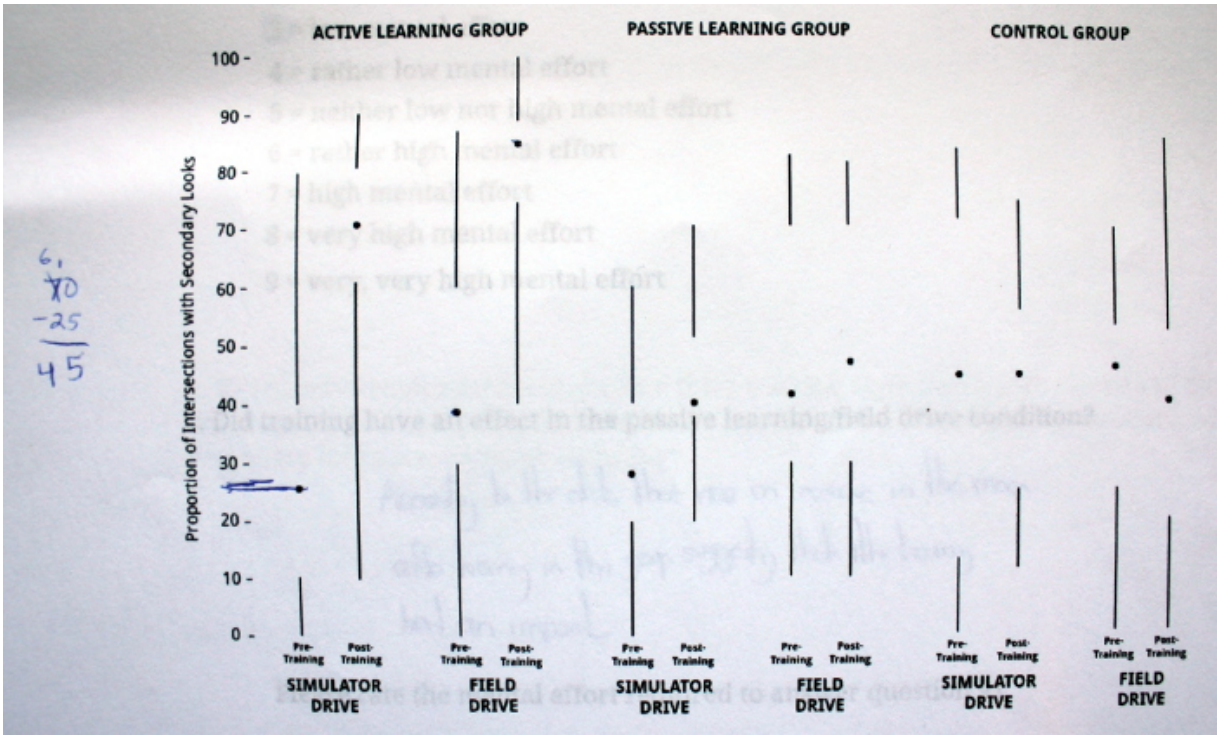

800
700
600
500
400
300
Reaction Time (ms)
Valid
Invalid
6
7
8
9
10
Age (years)

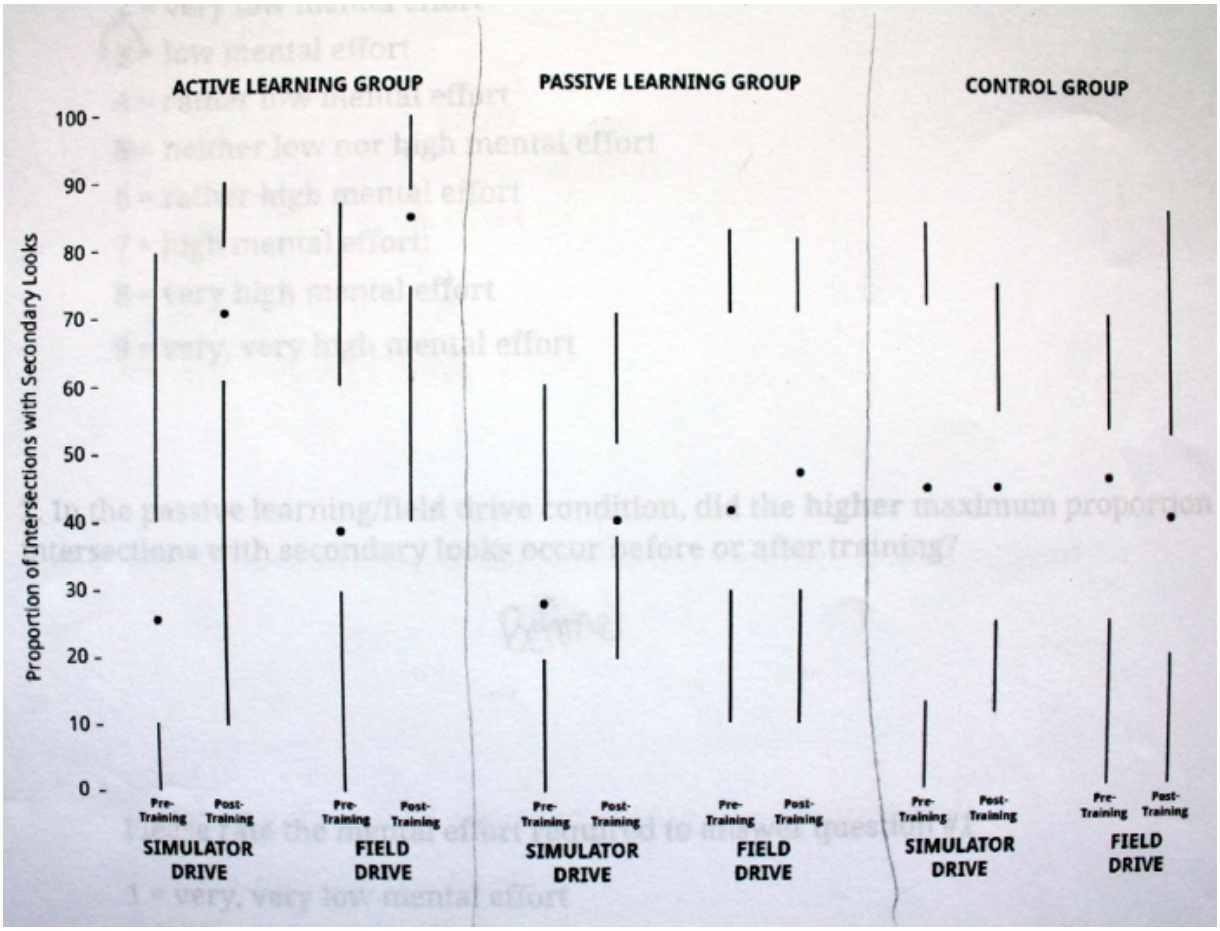

800
700
600
500
400
300
Reaction Time (ms)
Valid
Invalid
6
7
8
9
10
Age (years)

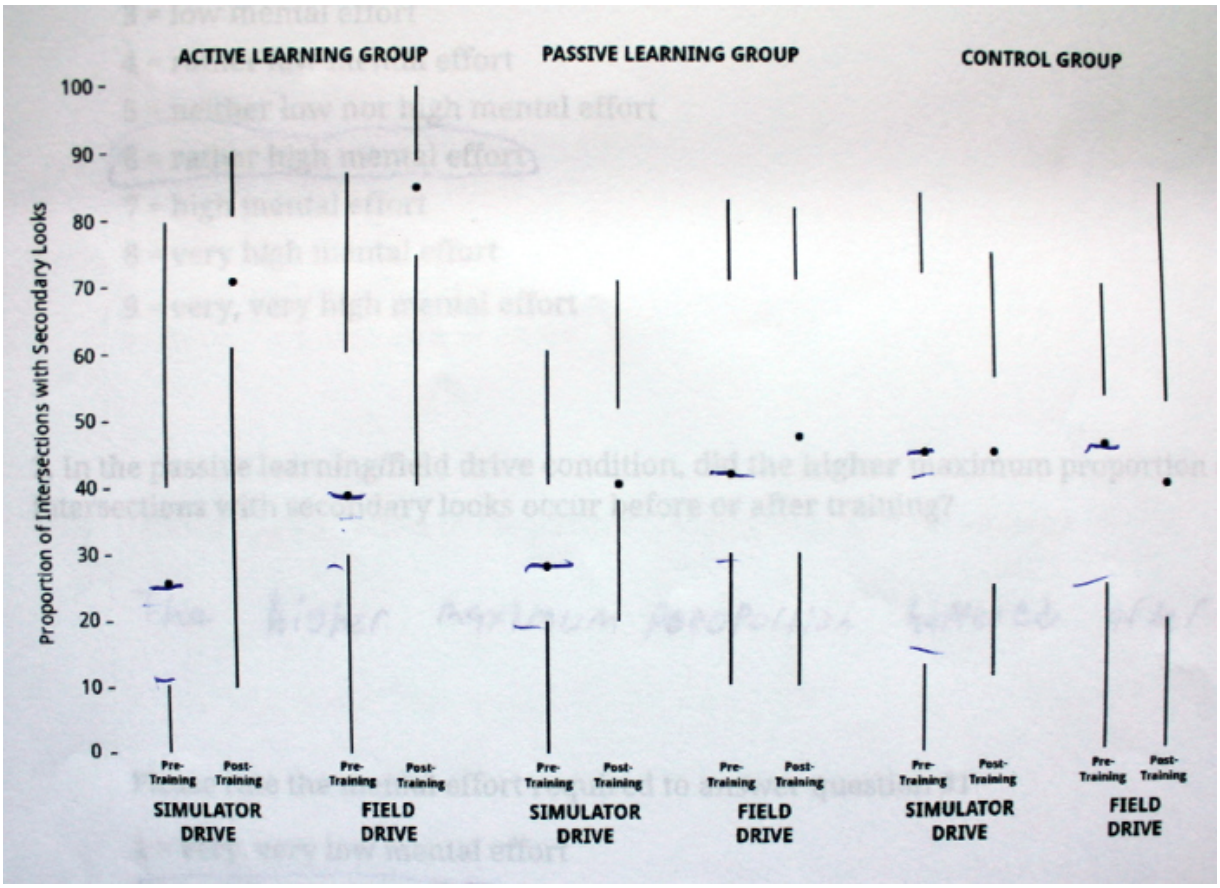

800
700
600
500
400
300
Reaction Time (ms)
Valid
Invalid
6
7
8
9
10
Age (years)

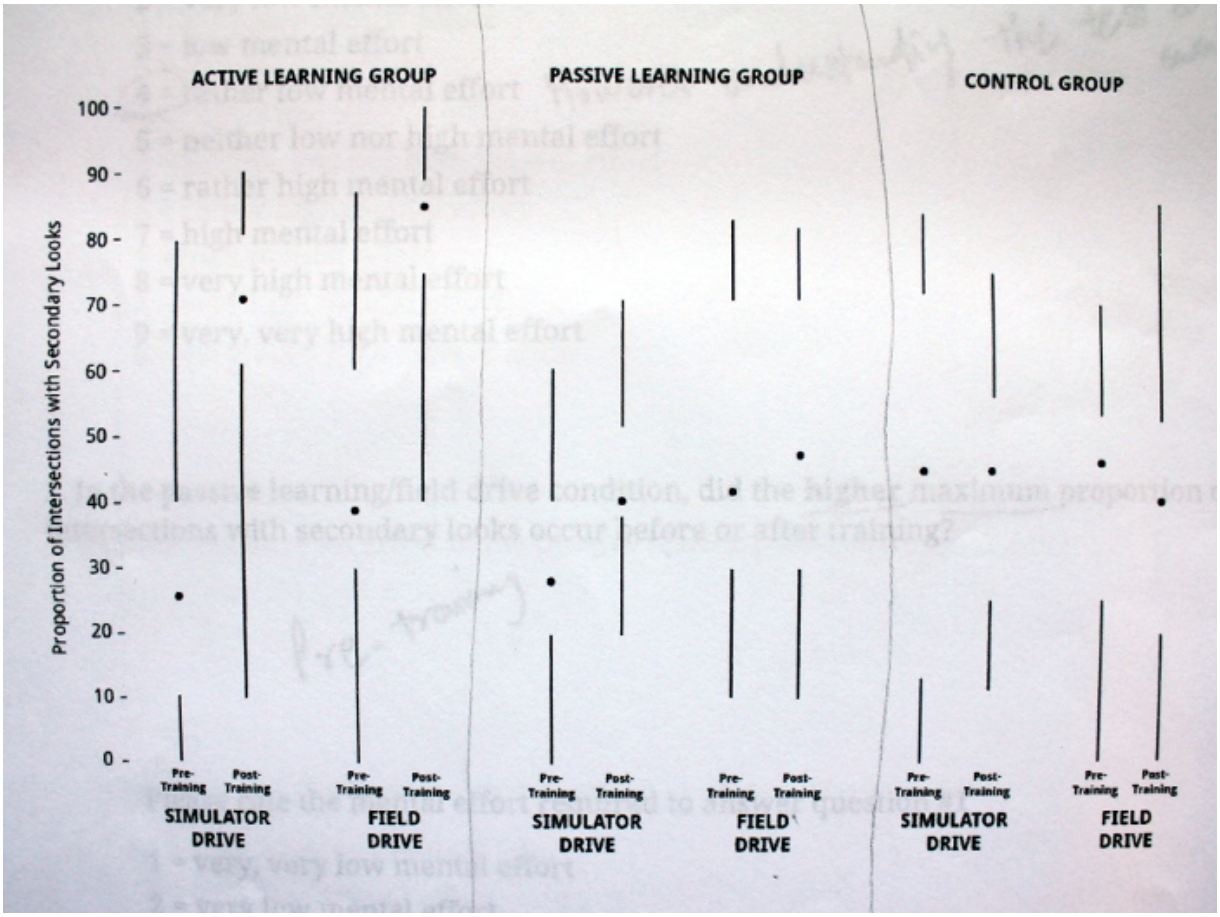

ACTIVE LEARNING GROUP
PASSIVE LEARNING GROUP
CONTROL GROUP
Proportion of Intersections with Secondary Looks
100
90
80
70
60
50
40
30
20
10
0
Pre-Training
Post-Training
SIMULATOR DRIVE
FIELD DRIVE
SIMULATOR DRIVE
FIELD DRIVE
SIMULATOR DRIVE
FIELD DRIVE

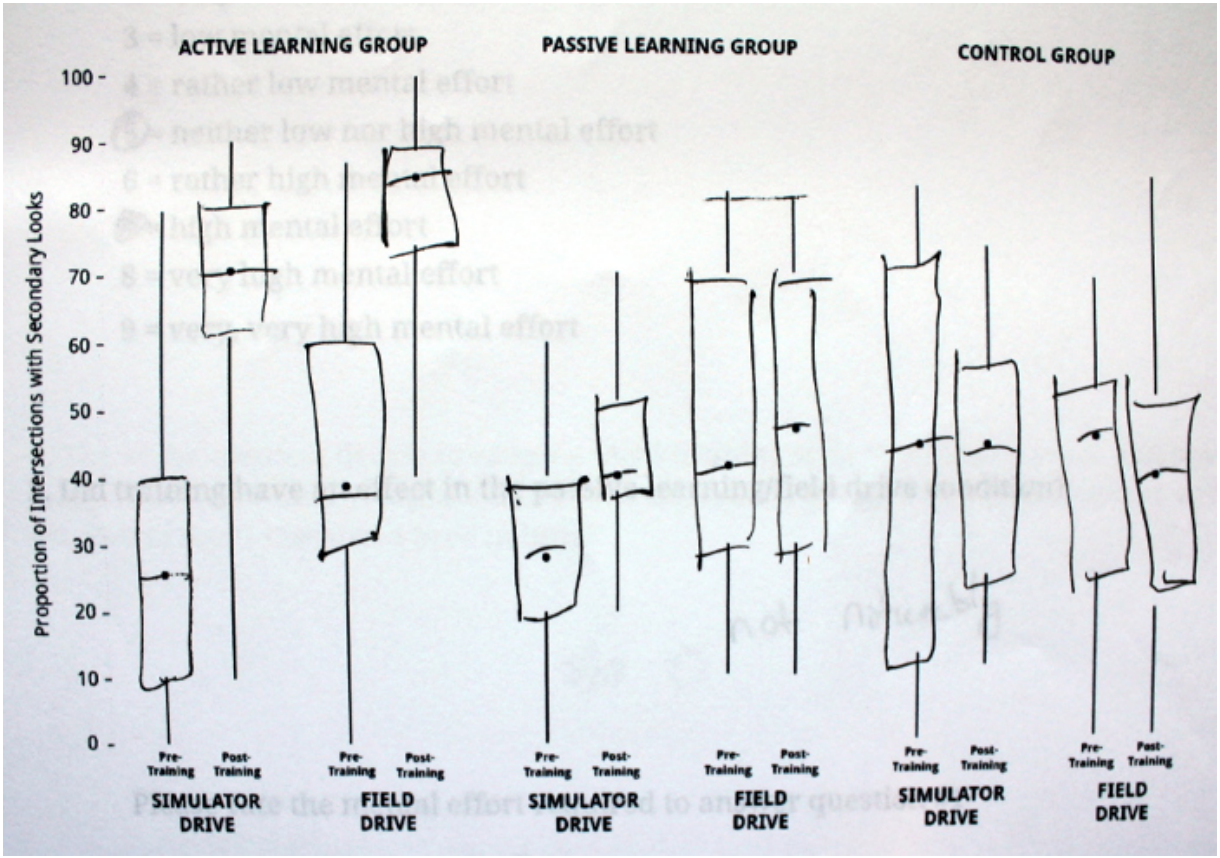

ACTIVE LEARNING GROUP
PASSIVE LEARNING GROUP
CONTROL GROUP
Proportion of Intersections with Secondary Looks
100
90
80
70
60
50
40
30
20
10
0
Pre-Training
Post-Training
SIMULATOR DRIVE
FIELD DRIVE


ACTIVE LEARNING GROUP
PASSIVE LEARNING GROUP
CONTROL GROUP
Proportion of Intersections with Secondary Looks
100
90
80
70
60
50
40
30
20
10
0
Pre-Training
Post-Training
SIMULATOR DRIVE
FIELD DRIVE


ACTIVE LEARNING GROUP
PASSIVE LEARNING GROUP
CONTROL GROUP
Proportion of Intersections with Secondary Looks
100
90
80
70
60
50
40
30
20
10
0
Pre-Training
Post-Training
SIMULATOR DRIVE
FIELD DRIVE


ACTIVE LEARNING GROUP
PASSIVE LEARNING GROUP
CONTROL GROUP
Proportion of Intersections with Secondary Looks
100
90
80
70
60
50
40
30
20
10
0
Pre-Training
Post-Training
SIMULATOR DRIVE
FIELD DRIVE


ACTIVE LEARNING GROUP
PASSIVE LEARNING GROUP
CONTROL GROUP
Proportion of Intersections with Secondary Looks
100
90
80
70
60
50
40
30
20
10
0
Pre-Training
Post-Training
SIMULATOR DRIVE
FIELD DRIVE


ACTIVE LEARNING GROUP
PASSIVE LEARNING GROUP
CONTROL GROUP
Proportion of Intersections with Secondary Looks
100
90
80
70
60
50
40
30
20
10
0
Pre-Training
Post-Training
SIMULATOR DRIVE
FIELD DRIVE

Appendix J

IRB Form C

R·I·T

Rochester Institute of Technology

RIT Institutional Review Board for the
Protection of Human Subjects in Research
141 Lomb Memorial Drive
Rochester. New York 14623-5604
Phone: 585-475-7673
Fax: 585-475-7990
Email: hmfsrs@rit.edu

Form C
IRB Decision Form

TO: Kevin McGurgan

FROM: RIT Institutional Review Board

DATE: **November 26, 2014**

RE: Decision of the RIT Institutional Review Board

Project Title – Data-ink ratio and task complexity in a graphical comprehension task

The Institutional Review Board (IRB) has taken the following action on your project named above.

☒ Exempt _46.101 (b) (2)_

Now that your project is approved, you may proceed as you described in the Form A.

You are required to submit to the IRB any:

- **Proposed** modifications and wait for approval before implementing them,
- Unanticipated risks, and
- Actual injury to human subjects.

Heather Foti, MPH
Associate Director
Office of Human Subjects Research