
Khalimahtul Akasah Idereena Khalil
University of Massachusetts Amherst

Follow this and additional works at: https://scholarworks.umass.edu/theses

https://doi.org/10.7275/7765832

This thesis is brought to you for free and open access by ScholarWorks@UMass Amherst. It has been accepted for inclusion in Masters Theses 1911 - February 2014 by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.
EXPERT-NOVICE DIFFERENCES:
VISUAL AND VERBAL RESPONSES IN A TWO-GROUP COMPARISON TASK

A Thesis Presented
by
KHALIMAHTUL IDEREENA AKASAH KHALIL

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

MASTER OF SCIENCE
February 2005
Psychology
EXPERT-NOVICE DIFFERENCES:

VISUAL AND VERBAL RESPONSES IN A TWO-GROUP COMPARISON TASK

A Thesis Presented

by

KHALIMAHTUL IDEREENA AKASAH KHALIL

Approved as to style and content by:

Clifford Konold, Chair

Carole Beal, Member

Alexander Pollatsek, Member

Melinda Novak, Department Head
Psychology
I would like to thank my advisor, Cliff Konold, for his never-ending patience, thoughtful guidance, and sponsorship throughout my time at UMass. I would like to thank members of my committee, Alexander Pollatsek and Carole Beal supporting this project. I would also like to express my sincerest gratitude to Arnold Well for spending his time with me on suggestions for the project, as well as conversations about life.

I would like to extend my thanks to Charles Clifton, Utama Abdul Wahid, and Michael Verde for technical assistance, to Izabela Santos and Bilal Ghandour for helping me recruit participants, and to Amitabh Bharadwaj and Michael Stroud for helping me read my manuscript. I would also like to extend my thanks to undergraduate and faculty members who sacrificed their time to participate in this project.

I wish to express my heartfelt thanks to all my friends near and far who always supported me and kept me in good spirits. I hope to see you all again soon. My deepest love and gratitude to my parents and sisters for everything they have done for me.
ABSTRACT

EXPERT-NOVICE DIFFERENCES:

VISUAL AND VERBAL RESPONSES IN A TWO-GROUP COMPARISON TASK

FEBRUARY 2005

KHALIMAHTUL IDEREENA AKASAH KHALIL, A.B., BOWDOIN COLLEGE
M.S., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Cliff Konold

This study investigated four hypotheses regarding how experts and novice data analyzers decide whether two groups are different based on graphs of the data:

1) Expert and novice data analyzers use different techniques to visually search data presented in graphs. The results showed that experts used more horizontal search movements to visually scan graphs along the horizontal x-axis of the independent variable while novices used more vertical search movements.

2) When asked to compare two groups of data in a graph, expert and novice data analyzers offer different verbal justifications for their answer. Experts were found to use mostly global comparison methods to explain their decisions about the data whereas novices frequently used local comparison methods.

3) Question wording can affect how novices justify their decisions about group differences. The study used questions phrased either in terms of which group was “more” or which group was “less.” In formulating their verbal justifications, novices were more likely overall to make exclusive reference to the data on the right side of the graphic display, regardless of question wording. However, the wording did appear to affect the
probability of novices making references to the right (with the “more” questions) or left (with the “less” questions), and this effect was in the predicted direction.

4) Question wording can affect where in the graph novices look to make their decision about group differences. The study showed that novices spent more time inspecting the side of the graph alluded to by the “more” or “less” wording of the question.

This study suggests that expert data analyzers use a visual search strategy that allows them to locate and use global features of the graphs, such as means or modal clumps. Novice data analyzers, on the other hand, use a visual search strategy that allows them to locate and use local features of the graph to compare the two groups using non-normative methods. More importantly, the study shows that the verbal justifications both experts and novices provide for why the groups are different or not are consistent with the ways they visually inspect the graphs to arrive at those decisions.
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>iii</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>iv</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>ix</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>x</td>
</tr>
<tr>
<td>1. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>Terminology</td>
<td>2</td>
</tr>
<tr>
<td>Research on Graph Comprehension</td>
<td>4</td>
</tr>
<tr>
<td>Contributions from Psychologists: Visually Searching Graphs</td>
<td>4</td>
</tr>
<tr>
<td>Contributions from Statistics Education Researchers: How People Talk about Data</td>
<td>6</td>
</tr>
<tr>
<td>Modal Clumps: An Informal Average to Summarize a Group</td>
<td>7</td>
</tr>
<tr>
<td>Two Approaches in Graph Comprehension: Local and Global</td>
<td>8</td>
</tr>
<tr>
<td>Combining Research Methods in Psychology and Statistics Education</td>
<td>9</td>
</tr>
<tr>
<td>Making Group Comparisons</td>
<td>13</td>
</tr>
<tr>
<td>Cutpoints: An Alternative Method of Comparing Two Groups</td>
<td>13</td>
</tr>
<tr>
<td>Slices: A Local Method of Comparing Two Groups</td>
<td>15</td>
</tr>
<tr>
<td>Choosing Slices and Cutpoints</td>
<td>18</td>
</tr>
<tr>
<td>Expert Descriptions of Data</td>
<td>19</td>
</tr>
<tr>
<td>2. EXPERIMENT OVERVIEW</td>
<td>21</td>
</tr>
<tr>
<td>General Task Overview and Goals</td>
<td>21</td>
</tr>
<tr>
<td>Predictions</td>
<td>23</td>
</tr>
<tr>
<td>Method</td>
<td>23</td>
</tr>
<tr>
<td>Participants</td>
<td>23</td>
</tr>
<tr>
<td>Materials</td>
<td>24</td>
</tr>
</tbody>
</table>
APPENDICES

A. DEMOGRAPHIC QUESTIONNAIRE ................................................................. 75
B. RFV CODE ................................................................................. 77
C. HOW TO CODE A TRANSCRIPT .................................................. 81
D. STACKED-DOTPLOTS OF HM DISTRIBUTIONS ...................... 86
E. STACKED-DOTPLOTS OF PROPTIMER DISTRIBUTIONS .......... 87

REFERENCES .................................................................................. 88
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Modified from Table 1, Ben-Zvi and Arcavi (2001).</td>
<td>11</td>
</tr>
<tr>
<td>2. Coding scheme used for coding participant justifications.</td>
<td>43</td>
</tr>
<tr>
<td>3. Table showing the number of novices who mentioned no tails, the LeftTail,</td>
<td>51</td>
</tr>
<tr>
<td>or the RightTail in response to a “less” or “more” worded question.</td>
<td></td>
</tr>
<tr>
<td>4. The number of LeftTail and RightTail responses mentioned in the 12 different</td>
<td>52</td>
</tr>
<tr>
<td>graph pairs given more- or less-worded problems.</td>
<td></td>
</tr>
<tr>
<td>5. Table 3, with the responses to the Numbness context removed, showing the</td>
<td>52</td>
</tr>
<tr>
<td>number of novices who mentioned no tails, the LeftTail, or the RightTail</td>
<td></td>
</tr>
<tr>
<td>in response to a “less” or “more” worded question.</td>
<td></td>
</tr>
<tr>
<td>6. Percentages for types of comparison methods used in the justification task by</td>
<td>55</td>
</tr>
<tr>
<td>expertise meandiff conditions.</td>
<td></td>
</tr>
<tr>
<td>7. Mean and confidence intervals for RT in the three meandiff conditions.</td>
<td>57</td>
</tr>
<tr>
<td>8. Mean proportion of time viewing the right side of the graph pair (PropTimeR)</td>
<td>61</td>
</tr>
<tr>
<td>and 95% confidence intervals given question wording and level of expertise.</td>
<td></td>
</tr>
<tr>
<td>9. Mean HM and PropTimeR (and their 95% confidence intervals) of experts and</td>
<td>63</td>
</tr>
<tr>
<td>novices.</td>
<td></td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. A graph pair showing the length 100 flowers from two species, Versicolor and Versinica</td>
<td>3</td>
</tr>
<tr>
<td>3. Modified from Table 1 in Konold, Robinson, Khalil, Pollatsek, Well, Wing, and Mayr (2002)</td>
<td>8</td>
</tr>
<tr>
<td>4. A choropleth graph (Figure 6, Ratwani et al., 2003) with a sample of eye-movement records showing where a participant looked in this graph while answering a global question.</td>
<td>10</td>
</tr>
<tr>
<td>5. Adapted from Figure 2, Ben-Zvi and Arcavi (2001)</td>
<td>12</td>
</tr>
<tr>
<td>6. Modified from Figure 13 in Lehrer, Schauble, Carpenter, and Penner (2000), showing the width of plants receiving two different doses (6 or 18 pellets) of fertilizer.</td>
<td>14</td>
</tr>
<tr>
<td>7. Two examples of how data can be “sliced” to make a local comparison between two groups.</td>
<td>15</td>
</tr>
<tr>
<td>8. Table and excerpt from Konold, Pollatsek, Well, and Gagnon (1997)</td>
<td>16</td>
</tr>
<tr>
<td>9. Pictorial representations of the slices students used according to excerpts from Watson and Moritz (1999).</td>
<td>17</td>
</tr>
<tr>
<td>10. Frequency Table for variables “Curfew” grouped by “License” and an excerpt from M and J comparing data in the table. From Konold et al., (1997).</td>
<td>19</td>
</tr>
<tr>
<td>11. The 12 graph pairs used in this experiment</td>
<td>25</td>
</tr>
<tr>
<td>12. An example of a graph pair without being obscured by the RFV</td>
<td>29</td>
</tr>
<tr>
<td>13. The same graph pair as shown in Figure 13 when viewed through the RFV viewing window.</td>
<td>29</td>
</tr>
<tr>
<td>14. A sample of the RFV output</td>
<td>32</td>
</tr>
<tr>
<td>15. Practice graph pair of hypothetical data</td>
<td>34</td>
</tr>
</tbody>
</table>
16. Three screenshots showing examples of what participants saw during each trial.

17. A searchpath from the novice participant 027 who, on the fish stimulus, used mostly horizontal movements (HM = 0.707).

18. A searchpath from the novice participant 105 who, with the soap stimulus, used mostly vertical movements (HM = 0.205).

19. The four quadrants of the "Fish" stimulus.

20. An example of a participant with low PropTimeR (.005).

21. The distribution of comparison methods for expert and novice participants.

22. The distribution of comparison methods for novice participants, as coded independently by two coders.

23. The distribution of comparison methods for expert participants, as coded independently by two coders.

24. The two graph pairs used in the study.

25. Novice and expert HM and 95% confidence intervals for the three meandiff conditions.

26. An example of a value bar graph showing the weight of students.
CHAPTER 1
INTRODUCTION

Data are ever present and routinely collected. Understanding the data that are presented in statistical graphs is fast becoming an essential skill in today’s society. Data analysis is now a topic introduced early in the United States mathematics curriculum (National Council of Teachers of Mathematics, 2000). Children in kindergarten carry out classroom surveys to determine favorite colors or food items (Russell, Schifter, & Bastable, 2002); middle school students in biology measure and compare plant growth rates in different environmental conditions (Lehrer, Schauble, Carpenter, & Penner, 2000).

Perhaps because of the push to introduce even young students to statistics and data analysis, there has been a rise in the number of statistics education researchers investigating how children, including those with some basic statistics instruction, interpret graphs (e.g., Ben-Zvi & Arcavi, 2001, Cobb, 1999; Konold & Higgins, 2003; Mokros & Russell, 1995). These researchers have reported that many students use non-normative techniques to describe data in graphs. However, being a novice data analyzer is not simply a matter of age. Even adults, such as nurses who interpret data daily (Noss, Pozzi, & Hoyles, 1999) and teachers who instruct middle and high-school students in statistics (Makar & Confrey, in press), make statements that differ from what statisticians would probably make.

This thesis investigated how novices and experts at data analysis structured their visual examinations of data in graphs, and how these visual searches were related to subsequent verbal justifications. The hope is that by better understanding how experts
and novices differ in how they think about and work with data will help us improve our instruction. Before presenting details of this study, I will clarify some terminology and review some of the relevant research on graph comprehension.

**Terminology**

There are a number of terms that I use in this thesis, which I introduce using the following narrative.

Imagine that you had a garden that you wanted to grow flowers in. You go to the local garden nursery to find out which iris plants grow bigger flowers. You measure the lengths of 50 Versinica and 50 Versicolor irises. Each flower you observe is a "data case," and the entire set of 100 flowers forms a "data set," or simply "data." You plot the data in a graph like the one below to compare the two groups so that you can decide which species grows bigger flowers (Figure 1). This type of graph, which I used as stimuli in this study, is called a "graph pair."

A graph pair is created using physical markings, in this instance, lines, circles, and text, arranged in a special configuration. Each physical marking, or "graph component," serves a purpose. Each circle represents a data case, i.e. a flower observed and measured at the nursery. The bold upper case text on the top left, "SPECIES," and bottom center, "LENGTH (mm)," indicate the two variables, or attributes, on which the flowers were measured. The two horizontal lines serve as axes for the "LENGTH" variable for the two groups. Smaller vertical lines along the bottom horizontal axis act as tick marks to create a numberline indicating the numeric scale of this variable.
Figure 1. A graph pair showing the length 100 flowers from two species, Versicolor and Verginica.

The particular way the graph components are configured is also purposeful. This graph pair is made from a pair of stacked-dotplots. In a stacked-dotplot, the position of a circle on the numberline indicates the length of the flower. Circles stacked above the top horizontal line represent the Versicolor flowers and those on the bottom horizontal line the Verginica flowers. For example, the two circles above the 50mm tick mark on the bottom horizontal axes indicate two Verginica flowers that were 50mm long.

In this flower example, you investigated how large the two types of irises grew. In such an investigative situation, you acted as a "data analyzer," carrying out a "data analysis" task. In this thesis, I use the term "data analyzers" to refer to people who examine data in graphs, or other representational forms, in an attempt to interpret and
make conclusions about those data. Data analyzers can range from complete novices to well-practiced experts.

I refer to the manner in which data analyzers search a graph as a "visual search pattern," "visual movement," or "searchpath." These terms are used interchangeably. I will also refer to two general patterns data analyzers use to visually scan data cases in a graph: using horizontal movements where they visually scan the data by moving their visual focus between the left and right sides of the graph, and vertical movements, where they scan the data cases up and down.

**Research On Graph Comprehension**

Psychologists and education researchers have used different approaches to explore how people comprehend graphs. In this section I first review what psychologists have found about how data analyzers look at graphs. Next I review what statistics education researchers have learned about how novices summarize groups of data and the two approaches these novices generally use when thinking about data. I then review research that further supports the notion that data analyzers use two general approaches to explore data. Fourth, I review two non-normative techniques that novice data analyzers use to make group comparisons. Finally, I review how expert data analyzers talk about data in graphs.

**Contributions from Psychologists: Visually Searching Graphs**

Psychologists have generally studied how data analyzers process graphs at the perceptual level. The questions they have pursued include measuring which graph components (lines, angles, text, etc.) people can judge most quickly as being the same or
a. Carpenter and Shah examined the average time data analysts spent looking at line graphs such as this.

b. The five regions that participants could focus on.

c. Arrows show the numerous transitions a typical data analyzer made between the five regions during a data analysis task. The direction of each arrow indicates the origin and target regions made during different transitions.

Figure 2. Diagrams from Carpenter and Shah (1998). These were modified from Figures 1, 4, and 5 in their paper.
different (e.g., Simkin & Hastie, 1987; Cleveland & McGill, 1984, 1985), how much time people spend examining different sections in the graph, such as the legend and axes (e.g., Carpenter & Shah, 1998), and how information about different graph components are visually processed (Pinker, 1990).

Carpenter and Shah (1998) characterized how undergraduates visually examined statistical graphs. They examined how data analyzers looked at line graphs (Figure 2a) showing data on three variables. For conducting their analyses, the researchers divided the line graphs into five different regions (Figure 2b) to examine the nature of participants’ eye movements over these areas as a function of the complexity of the data in the “Pattern” region.

The authors found that data analyzers spent most of their time looking at the “Pattern” region of the graph, but made numerous transitions back and forth between the five regions (Figure 2c). As the data in the Pattern region became more complex (i.e., included more lines and more instances of line-crossovers) the number of these transitions increased.

Contributions from Statistics Education Researchers: How People Talk about Data

Motivated by the goal to improve instruction, statistics education researchers have generally focused on how students, who are usually novice data analyzers, verbally discuss data. The goal has been to gain insight into how students understand the data and their task. These researchers have studied how students make graphs (e.g., Lehrer & Schauble, 2000), how they summarize data (e.g., Konold, Robinson, Khalil, Pollatsek, Well, Wing, & Mayr, 2002), and how they compare two groups of data (e.g., Gal,
Rothschild, & Wagner, 1989; Watson & Moritz, 1999). These researchers have reported that many students use informal, and sometimes non-normative, techniques to describe data in graphs. For example, Konold et al. (2002) found that many students used a “modal clump” instead of an average to describe the central tendency of a group.

Modal Clumps: An Informal Average to Summarize a Group

Konold, Robinson, Khalil, Pollatsek, Well, Wing, and Mayr (2002) examined how students summarized data in stacked-dotplots. Seventh and ninth graders examined stacked-dotplots of data they collected as part of the Roadkill project. In the Roadkill project, students counted and classified dead animals they observed along town roadways over several weeks. They shared their data with students in other classrooms over the Internet.

Konold et al. interviewed students in pairs or teams of four. The interview questions were based on the data students had collected. At the start of each interview, before the students saw any graphs, the interviewer asked the students to estimate, “How many dead animals do you tend to see each day?” Students tended to use a range to provide an estimate of the data. For example, one team, R2, said:

Grant : Well, it would change every day, depending on if it was raining, or sunny...
Ryan : Or like weather, climate, stuff like that.
Interviewer : Right.
Anita : But on average, probably be like 1 or 2. Not many.

The interviewer then asked the students to generate a stacked-dotplot of made up data showing the distribution of dead animals they might expect to see on the roadways over a 15-day period. Most of the graphs generated were unimodal and roughly
symmetrical, like the one in Figure 3. Konold et al. then asked the students, “How would you summarize this plot?” The students tended to summarize and describe the data using a modal clump. For example, Pat, a member of team D4, summarized the data in Figure 3 as, “Um, this, it's not, it's not like too many. It's not more around 12's, but they're mostly in the range of the middle numbers: 4 through 8.”

![Figure 3](https://edutel.musenet.org:8042/roadkill/)

**Figure 3.** Modified from Table 1 in Konold, Robinson, Khalil, Pollatsek, Well, Wing, and Mayr (2002).

Konold et al. suggested that the students were using the modal clump in a way that was akin to the InterQuartile Range in that it indicated the location of the center bulk (around 50%) of the data. The notion of the modal clump built on recent research showing how people often used ranges to point to the location of the central majority of the distribution (e.g., Bakker, 2001; Cobb, 1999; Konold & Higgins, 2003).

**Two Approaches in Graph Comprehension: Local and Global**

Several researchers have characterized how data analyzers reason about data as either “local” or “global” (e.g., Bakker, 2001; Ratwani, Trafton, & Boehm-Davis, 2003;

---

1 http://edutel.musenet.org:8042/roadkill/
Konold & Higgins, 2003; Mokros & Russell, 1995). A global approach entails describing data in terms of the emergent properties of the whole data set, such as the mean of a distribution, a line of best fit in a scatterplot, or the percentage of cases beyond a certain value. A local approach entails describing data by focusing on specific cases or subsets of data in a graph, without relating those cases to the entire data set as whole.

Ratwani, Trafton, and Boehm-Davis (2003) used the local-global distinction to describe the nature of both the data, and the questions asked about data. They defined local data as information that was readily available and easily read off of the graph; local questions asked data analyzers to obtain such local data from the graph. An example of a local question about the data in Figure 1 is “What is the length of the longest flower?” Alternatively, global data requires data analyzers to abstract general information from the data, while global questions asked for such global data. An example of a global question concerning Figure 1 is “What is the mean of the Versicolor flowers?”

Combining Research Methods in Psychology and Statistics Education

Ratwani, Trafton, and Boehm-Davis (2003) combined the two tracks of research (i.e., statistics education and psychology) to understand graph comprehension. To test whether global and local questions elicited global and local responses, respectively, Ratwani et al. (2003) presented psychology undergraduates choropleth graphs of the population sizes in different state counties (Figure 4) and asked two types of questions: “local” questions, such as “What is the population of Victorville county?”; and “global” ones, such as, “What is the general trend of population growth in this graph?” The researchers looked both at how participants verbally answered these questions and at how participants examined the graphs, the latter by monitoring their eye movements.
They found a strong relationship between what students said and where they looked in the graph as they answered questions about data. Most students answered local questions with local answers^2, and spent more time examining labels in the graph. Global questions tended to be answered with global responses^3, and in these instances participants spent more time visually examining the borders of counties in the map.

![Figure 4](image-url)

**Figure 4.** A choropleth graph (Figure 6, Ratwani et al., 2003) with a sample of eye-movement records showing where a participant looked in this graph while answering a global question. In this graph, the color of a county indicates population size.

---

^2 An example of a response is “The population of Victorville county is 20,451 to 35,622.”

^3 An example of a response is “There is more blue [population 30,000-40,000] on the graph, and less orange [10,000-20,000].”
Ratwani, Trafton, and Boehm-Davis’ (2003) result is not as trite as it may seem. Other researchers (e.g., Konold et al., 2002; Ben-Zvi & Arcavi, 2001) have reported instances where, with a single question, students alternated between local and global descriptions of the data.

Ben-Zvi (1999) and Ben-Zvi and Arcavi (2001) described two students, A and D, who wrote summaries that emphasized local features of the data despite the fact that the two students had just discussed with the teacher various global features of the data set. The data A and D were considering showed the Olympic gold medal times of the men’s 100-meter race. The data were presented first in a table (Table 1) and then in a graph (Figure 5). The question they were asked was “What do we learn from this table?”

Although the question is neither clearly global nor local, A and D responded to it by giving local responses, at least initially:

We don’t learn anything special. There is nothing special here! For example, the record time here is smaller; here it’s bigger.

and,

A row describes when the Olympiad took place, in what place, the winning athlete’s name...

<table>
<thead>
<tr>
<th>Year</th>
<th>City</th>
<th>Athlete’s name</th>
<th>Country</th>
<th>Men’s 100m Olympic time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1896</td>
<td>Athens</td>
<td>Thomas Burke</td>
<td>USA</td>
<td>12.0</td>
</tr>
<tr>
<td>1900</td>
<td>Paris</td>
<td>Francis Jarvie</td>
<td>USA</td>
<td>10.8</td>
</tr>
<tr>
<td>1904</td>
<td>St. Louis</td>
<td>Archie Hahn</td>
<td>USA</td>
<td>11.0</td>
</tr>
<tr>
<td>1908</td>
<td>London</td>
<td>Reginald Walker</td>
<td>South Africa</td>
<td>10.8</td>
</tr>
<tr>
<td>1912</td>
<td>Stockholm</td>
<td>Ralph Craig</td>
<td>USA</td>
<td>10.8</td>
</tr>
<tr>
<td>1920</td>
<td>Antwerp</td>
<td>Charles Paddock</td>
<td>USA</td>
<td>10.8</td>
</tr>
<tr>
<td>1924</td>
<td>Paris</td>
<td>Harold Abrahams</td>
<td>UK</td>
<td>10.6</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>1996</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Table 1. Modified from Table 1, Ben-Zvi and Arcavi (2001). This is a subset of the Olympic data that students A and D were analyzing.
Figure 5. Adapted from Figure 2, Ben-Zvi and Arcavi (2001).

After a few minutes, the students moved to the next part of the assignment, which instructed them to make a time-series plot of the data (Figure 5). Regarding this graph, the assignment asked, “What do we learn from this graph?” Again, the students first gave local descriptions of the data. However, with teacher intervention, they were eventually able to give global answers. For example, their initial local response was, “[We learn] what running time was achieved in what year.” After receiving assistance from the teacher, they responded,

A: Hold on for a second! And we also learn that the results don’t always increase.

... D: Yes, right. The records do not always increase.

... A&D: No! Here it increases, here it decreases, here it increases, the same, the same, increases, decreases,... [They continue for a while]

Eventually, with teacher encouragement, they began offering global statements such as, “[the data showed an] Improvement in records” and a “Decline in running times.”
At the end of the session, A and D wrote a summary that described both global and local features of the data, “The overall direction is increase in the records, yet there were occasionally lower (slower) results, than the ones achieved in previous Olympiads [e.g., comparing the winning time achieved in 1984 versus 1980].”

Making Group Comparisons: Two Non-normative Techniques

A common way for experts to compare two groups is to compare the location of the two averages, such as the means. I refer to this comparison method as “global displacement” because it describes the relative location of corresponding aggregate features of the groups. In addition to the centers (i.e., means, medians, modal clumps), data analysts could also look for displacements in the range of the data.

Returning to the students in the Roadkill project, Konold, Robinson, Khalil, Pollatsek, Well, Wing, and Mayr (2002) noted that even though the students summarized the data using modal clumps, very few students used the global displacement method to compare the modal clumps of the two stacked-dotplots. If students did not compare groups by comparing the displacement of aggregates, then what did they do?

Cutpoints: An Alternative Method of Comparing Two Groups

Several studies have found students using a cutpoint comparison method to compare two groups. This involves partitioning the data into two subsections using a value on the dependent variable and then comparing the numbers or percentages of cases in the two groups above or below this cutpoint. Lehrer, Schauble, Carpenter, and Penner (2000) observed third-graders investigating how two different amounts of fertilizer affected the growth of plants. The students measured the widths of the plants as an indicator of growth and plotted the data in a stacked-dotplot (Figure 6).
Figure 6. Modified from Figure 13 in Lehrer, Schauble, Carpenter, and Penner (2000), showing the width of plants receiving two different doses (6 or 18 pellets) of fertilizer.

One student concluded that plants given the 18-pellet dose grew more,

Peter: Because only 4 of the 12, only 1/3 of the 18-pellet plants are less than the widest 6-pellet plants

Peter argued that 18-pellet plants grew more by placing a cutpoint at “the widest 6-pellet plant,” which is at about a width of 125 mm. He then compared the proportion of 18-pellet plants (33% of the data) to the proportion of 6-pellet plants (100%) below the cutpoint. Peter used the cutpoint method in a global way because he compared the proportions of the two groups below 125 mm.

Data could also be compared using cutpoints in a local way. Suppose that another student, Chris, made the following argument based on the same data in Figure 6: the plants getting the 18-pellet does grew wider because there are 8 of them wider than 120 mm compared to only 1 of the 6-pellet dose. This would be a local comparison method
because the cases greater than 120 are isolated and their absolute counts compared with each other. Watson & Moritz (1999) reported students using cutpoints in the local way. Indeed, reasoning about data using absolute and not relative frequencies can sometimes be difficult for students (Bakker 2004), which prompted researchers to explore ways to develop proportional reasoning in students (e.g., Bakker, 2001).

**Slices: A Local Method of Comparing Two Groups**

In their study of high school students, Konold, Pollatsek, Well, and Gagnon (1997) reported that many students used "slices" to compare two groups, as exemplified in Figure 7. A slice may contain just one stack of cases (Figure 7) or several contiguous stacks of cases (Figure 7). In a graph pair, this method involves "slicing" across the two groups with two (Figure 7a) or more (Figure 7) dividing lines, and then comparing within that slice (or those slices) the number of cases in each group. Comparing the groups in this way is a local method because those cases are discussed in isolation, without relating them to the entire data set (e.g., by describing them proportionally).

**Figure 7** Two examples of how data can be "sliced" to make a local comparison between two groups.
In Konold, Pollatsek, Well, and Gagnon (1997), the high-schoolers used slices to compare data presented in 2-way frequency tables. Two students, P and R, investigated whether having a curfew was related to hours spent doing homework. On lines 275 and 277 in the excerpt below (Figure 8), P looked at slices at 12 and 14 homework hours and compared the number of students in the “with” and “without” curfew groups in those slices.

<table>
<thead>
<tr>
<th>Curfew</th>
<th>Hw</th>
<th>12</th>
<th>14</th>
<th>15</th>
<th>27</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>no</td>
<td>7</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>yes</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>total</td>
<td>9</td>
<td>4</td>
<td>7</td>
<td>9</td>
<td>1</td>
<td>15</td>
</tr>
</tbody>
</table>

Figure 8. Table and excerpt from Konold, Pollatsek, Well, and Gagnon (1997).

Watson and Moritz (1999) also reported students who compared graph pairs using slices. A 6th grader concluded that the number of facts recalled by the Yellow and Brown classes depicted in Figure 9 were the same because,

... the Brown class, they had kind of more people [i.e., three versus two] in the 6 and 7 [slice] in the higher scores (Figure 9a), but these [Yellow] people
had a lot more people [i.e., five versus three] in the middle [slice at 5], which kind of added up (Figure 9b), so they’re even... (p.155).

![Diagram](image)

**Figure 9.** Pictorial representations of the slices students used according to excerpts from Watson and Moritz (1999).

Slicing data may serve two purposes. First, data analyzers may feel that slicing makes the data more manageable, especially when the cases are numerous. Slicing creates smaller, more homogenous, sub-groups of data to work with. Creating these more manageable subgroups may "solve" the problem of variability in the data, allowing the person to compare subgroups in which the values of the dependent variable of interest do not vary much, if at all.

Second, data analyzers may use slices to locate those cases that they believe are most relevant for answering their question. For example, R and P above may have chosen the particular slice they did because they believed their question was asking only about students who spent considerable time studying. Indeed, I describe below another pair of students in the same study who were quite explicit about ignoring half of the relevant
data in a two-way table they were looking at because they thought those data were not pertinent to their question.

Choosing Slices and Cutpoints

When making comparisons between groups, how do data analyzers decide where to place slices or cutpoints? As indicated above, one possibility is that they base their decision on how the question is worded. It is often possible to incorporate key words or phrases in the question into the reply. For example, consider the question asked in the Watson and Moritz (1999) study, “Do Yellow class students recall *more* facts than Brown class students?” The answer to this question may be formulated as, “Yes, Yellow class students recall *more* facts because…” Likewise, “Do Yellow class students recall *fewer* facts than Brown class students?” may be answered with, “Yes, Yellow class students recall *fewer* facts because…” Question wording may also prompt data analyzers to incorporate into their responses cases from the corresponding end of the graph pair – the upper end in the case of “more” questions and lower end of the graph in response to “less” questions.

Another possibility is that data analyzers select a slice or cutpoint towards the right side of the axis, for example, because they misinterpret a “more” question as asking only about greater-valued cases. For example, two students, M and J, wanted to find out whether having a driver’s license was related to having a curfew (Figure 10, from Konold, Pollatsek, Well, & Gagnon, 1997). From lines 358 and 359, it seems that M and J believed that this was a question only about students with curfews; they stated that it was only appropriate to look at students with curfews and not those without curfews.
Table 4

Frequency Table for Variables “Curfew” Grouped by “License”

<table>
<thead>
<tr>
<th>License</th>
<th>Curfew</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>no</td>
<td>17 (0.31)</td>
<td>37 (0.69)</td>
</tr>
<tr>
<td>yes</td>
<td>34 (0.35)</td>
<td>64 (0.65)</td>
</tr>
<tr>
<td>total</td>
<td>51 (0.34)</td>
<td>101 (0.66)</td>
</tr>
</tbody>
</table>

355. I: Okay, and again, in that last comparison, we sort of ignored these two numbers [pointing to “no” curfew column]. Is that, is that all right to do?
356. M: Yeah, for what we are comparing.
357. I: And, how come?
358. J and M: We were looking at
359. J: license with [curfew], and not without. And you could do another question without, because our main thing was the license.
360. M: And the curfew.
361. I: Okay. So, if I asked you, “Are you less likely to have a curfew if you have a license vs. if you don’t have a license?” then you’d look at those other numbers?

Figure 10. Frequency Table for variables “Curfew” grouped by “License” and an excerpt from M and J comparing data in the table. From Konold et al., (1997).

These students effectively reinterpreted a global question (Which group of students is more likely to have a license?) as a local one (Of students who had a curfew, did more of them have a license or not have a license?).

Expert Descriptions of Data

To this point I have been describing mostly how novice data analyzers discuss data. Indeed most research by statistics educators have studied novices. In a series of studies, Trafton and his colleagues have begun examining expert data analyzers (e.g., Trickett, Fu, Schunn, & Trafton, 2000; Trafton, Kirschenbaum, Tsui, Miyamoto, Ballas, & Raymond, 2000; Trafton & Trickett, 2001). Trickett, Fu, Schunn, and Trafton (2000)
asked experts (e.g., astronomers, physicists, neuropsychologists, and meteorologists) to "think aloud" as they interpreted graphs as part of their everyday research program.

They found that when the patterns of data were consistent with the experts' expectations, those experts described them using formal scientific terminology. To describe anomalous data, however, the experts used perceptual terminology. For example, a neuropsychologist described brain activity data that were consistent with his expectations quite formally: “So there is the subcortical activation that is probably caudate.” In contrast, when graphs showed unexpected data, experts used perceptual terms such as “blob,” “bulge,” or “dipsy-doodle,” sometimes indicating its location in the graph (e.g., “lower right,” “northwest”).
CHAPTER 2

EXPERIMENT OVERVIEW

The primary purpose of this study was to explore the relationship between 1) how expert and novice data analyzers visually scanned graph pairs to reach a decision about group differences and 2) how they justified their decisions. As I have described above, Ratwani et al. (2003) examined where data analyzers looked when given global vs. local questions. The pattern of eye movements among different regions in a graph differed depending on whether the questions were about local or global characteristics of the data. In this thesis, I attempted to characterize not only where but how data analyzers visually searched data in graph pairs. I expected that data analyzers who conceived of their task at the global level would structure their visual search so as to locate global features of the data, and would also discuss the data globally. Those who conceived of the task at the local level would justify their decisions about group difference using local descriptors and would visually search graphs in ways that were consistent with this local view of data. I elaborate below the predicted nature of these locally-driven search techniques.

General Task Overview and Goals

Participants' first task was to view a series of graph pairs (e.g., Figure 1) to decide whether one of the groups tended to have higher or lower values than the other group. Their second task was to give a brief verbal response to justify the reason for their decision about the two groups. The participants included both experts and novices. I predicted that experts would use primarily aggregate methods to compare the two groups. Accordingly, in visually searching the graph pairs, I predicted that they would tend to
scan horizontally across the graph, locating the approximate average of each group so that they could compare them. When asked to justify their decisions, I also expected that they would describe a global method. Based on prior research, I expected that novices would be more likely to use local comparison methods, such as slices, to justify their decisions and that in visually searching the graphs they would tend to use more vertical movements, which locating and using cutpoints and slices would seem to require.

The secondary goal of this study was to explore whether question wording affected the way novices compared two groups of data. Recall that Konold, Pollatsek, Gagnon, and Well (1997) suspected that some students they interviewed used slices and cutpoints because they misinterpreted the question as asking about only subsections of the data. If this is the case, then I ought to be able to influence where novices look in a graph, and where they locate cut points and slices, by manipulating the wording of the question.

To investigate this possibility, the questions I gave participants about whether two groups differed were worded in two ways; in terms of which group was "more," or which group was "less." I predicted that if participants assumed "more"-worded questions were best answered using data in the right hand side of the graph pair, then in their justification they would be more likely to mention data in the right-tails of the distribution. Alternatively, a "less"-worded question would elicit more verbal justifications that mentioned data from the left hand side. I also predicted that novices would spend proportionally more time visually inspecting the side of the graph pair alluded to by the question wording – the right hand side for "more" questions, and the left hand side for "less" questions.
Predictions

To summarize, the main predictions in this study were: (1) Experts would tend to visually scan the graph pairs using horizontal movements while novices would scan the data using comparatively more vertical movements. (2) Experts would tend to justify their decision using global methods of comparison while novices would use local methods. For novices, whether questions were worded in terms of “more” or “less” would influence (3) the parts of the data they mentioned in their verbal justifications and (4) the percent of time they spent inspecting the corresponding side of the graph. I predicted that question wording would have no effect on expert performance.

Method

Participants

Thirty undergraduates (26 women) taking classes in the psychology department at the University of Massachusetts, Amherst were recruited as novice data analyzers and were compensated for their time with credits towards their psychology courses. Eleven were sophomores, 14 were juniors, and 5 were seniors. Their mean age was 20.0 years (sd = 1.4). They had taken a mean of 1.7 statistics courses and 2.0 mathematics courses.

Ten faculty members (2 women) in the departments of psychology and of statistics at the same university volunteered their time as expert data analyzers. Eight regularly taught statistics courses. Their mean age was 52.6 years (sd = 12.1). As undergraduates or graduate students, they had taken a mean of 5.5 statistics courses and 3.8 mathematics courses.
Materials

The stimuli were 12 graph pairs (Figure 12), each presented with a unique context (e.g., weight of fish, length of flowers, etc.). The contexts for the graph pairs were chosen such that participants would not have specific expectations either about whether the two groups would differ or how they might differ. Several graph pairs showed authentic data and contexts obtained from on-line databases (e.g., “flowers” from the Fisher’s Irises dataset from a university website\(^4\)). But most of the data and contexts were contrived (e.g., “soap,” “trees”).

The data in each graph pair satisfied several criteria. The two groups had the same or similar ranges, had similar standard deviations and shape (e.g., fairly normal, somewhat skewed, or somewhat flat), and were mostly singly peaked. The two groups varied in terms of the difference between their means. These differences ranged from small, to medium, to large (see Figure 11), as defined by their “mean difference ratio” (meandiff):

\[
\text{Mean difference ratio} = \frac{|X_1 - X_2|}{X_{sd}}
\]

\(X_1\) = mean of the top group in the graph pair
\(X_2\) = mean of the bottom group in the graph pair
\(X_{sd}\) = mean of the standard deviations of the two distributions

\(^4\) [http://w\v\v-unix.oit.umass.edu/~statdata/statdjta/data/iris.txt](http://w\v\v-unix.oit.umass.edu/~statdata/statdjta/data/iris.txt) for a description of the data set and [http://www-unix.oit.umass.edu/~statdata/statdata/data/iris.dat](http://www-unix.oit.umass.edu/~statdata/statdata/data/iris.dat) for the data
Figure 11. The 12 graph pairs used in this experiment. Graph pairs in the first three rows have continuous x-axes variables while the bottom row of graph pairs have integer x-axis variables.
Figure 11, continued

Medium difference between the means: Meandiff ratio ≈ 0.65

**FISH**

unmod

genmod

**TOWN**

Ganaaq

Dundas

**TOP**

Top A

**DAY**

cold

**Spin-Time (secs)**

Top B

**Hot**

NUMBER OF ROADKILL

Continued, next page
Figure 11, continued

Large difference between the means: Meandiff ratio ≈ 1.90

**SPECIES**
- Versicolor

**SUBSTANCE**
- Copper

**DRUG**
- Pandol

**TREATMENT FORMULA**
- Q23

**TIME FOR NUMBNESS TO BEGIN (mins)**

**NUMBER OF SPROUTS**

**LENGTH (mm)**

**TWISTABILITY (γ)**
The intent of the large meandiff stimuli was to make it obvious that the two groups differed. Similarly the intent of the small meandiff was to clearly suggest that the two groups, though not identical, were basically the same. The medium group served as an intermediate level between the small and large meandiff stimuli.

The stimuli were presented on a plain white background on either a 17” or 19” computer monitor. The total area of the graph stimulus was approximately 11” wide by 7” high (i.e., 792 x 477 pixels). Each group in a graph pair contained 50 data cases. Each data case was indicated in the graph with a circle. Half the stimuli showed the mean of the top group positioned to the right of the mean of the bottom group, and half showed it to the left.

Apparatus

An audio-tape recorder was used to record participants’ verbal responses. A computer that ran the Restricted Focus Viewer (RFV) program was used both to structure and to record how participants visually examined the graph pairs. The RFV is a software program developed by Blackwell, Jansen, and Marriott (2000) as an alternative to standard eye-tracking equipment (see Appendix A for RFV code).

I set up the RFV software to restrict the view of the data cases in a graph pair to a square viewing window of approximately 2 x 2 inches (i.e., 150 x 150 pixels). Participants controlled what appeared in the viewing window by moving the window over the screen using the computer’s mouse. Labels, axes, and scales on the graph pairs were always visible (see Figures 12 and 13).

---

5 1 inch = 72 pixels
Figure 12. An example of a graph pair without being obscured by the RFV

Figure 13. The same graph pair as shown in Figure 13 when viewed through the RFV viewing window.

The RFV recorded the elapsed time (in ms) and location (x- and y-pixel coordinates) of the center of the viewing window about every 22ms (sd=3) during the visual search task. At each observation, the RFV also recorded whether the content of the viewing window was visible or not. The viewing window was set to obscure the data.
within the viewing window if participants moved the viewing window too quickly across the computer screen (i.e., faster than about 7000 pixels (10") per second). The analysis in this thesis only used observations for which the content of the viewing window was visible and the viewing window was actually on the 11”x7” area of the graph pair itself (Figure 12) as opposed to being in the background areas of the display that surrounded the graph pair. These background areas contained the text that described the context and the questions for the trial (see Figure 15).

Some tasks asked participants to make responses about the data on the computer screen, and these responses were also recorded by the RFV. The RFV produced output files that displayed the movement of the viewing window during the experiment, which included a timestamp and the x- and y-coordinates of the center of the viewing window (Figure 14). The RFV required a search procedure that no doubt was different from what participants would use if the contents of the graphs were never obscured. There were two reasons for using the RFV rather than a conventional eye-tracking apparatus.

First, using a restricted view helped structure the way participants gathered visual information from data. If participants in this study were allowed to see the graph pairs in their entirety, they (and in particular experts) may have been able to assess group differences with just a few fixations; thus no clear pattern would have been discernable in their searchpaths (Liversedge, 2003, personal communication in reference to Feeney, Hola, Liversedge, Findlay, & Metcalf, 2000, follow-up study). By forcing participants to process the graph pairs more sequentially rather than holistically, I hoped to induce more structured sequential search patterns. This was critical for establishing whether
participants sought information from graphic displays using more horizontal or vertical search movements.

Second, using a restricted viewing window controlled the size of the visible stimulus. This was important because research suggests that with visual information related to their expertise, experts are able to see and process larger areas than are novices. For example, Reingold, Charness, Pomplun and Stampe (2001) found that chess experts, compared to chess novices, had larger “visual spans,” which allowed them to process meaningful patterns of chess configurations more quickly.

One limitation of the RFV is that we cannot be certain that participants were, at any given time, actually looking at the contents of the RFV viewing window. For example, a participant may have been viewing the numberline on the bottom axis while the viewing window was located on the top stacked-dotplot. Nevertheless, I assumed that when data were visible through the viewing window, participants would be looking at them and processing the information. This assumption is similar to the assumption made in conventional eye-tracking studies of reading whereby it is a “general belief that eye fixation patterns … tell us something about perceptual and cognitive processes” (Loftus, 1983, p359) by giving a “word-by-word indication of processing” (McConkie, 1983, p.91). This assumption is made even though researchers acknowledge that where the eyes are currently fixated may not always be indicative of what the person is mentally processing (McConkie, 1983).
RFV Version 2.1 Output Data File
Subject:           Novice113
Input Filename:    exptmscba.rfv
Date:              Sep 30, 2002

Component:        termites
                  0  1094  524  663
                  0  1109  523  660
                  0  1125  522  658

                  1  20297  649  466  B
                  1  20312  651  466  F
                  1  20328  653  466  F
                  1  20344  654  466  F
                  1  20375  655  466  F

                  0  31344  682  660
                  0  31453  683  660

Response:         NO, they are about the same
Time (ms):        32109

Figure 14. A sample of the RFV output. The row labeled "Component:" indicates which graph pair was shown to the participant, in this case the termites problem. The ellipses indicate portions of the output that have been omitted.

The leftmost column of numbers indicates whether the mouse was located on (1) or off (0) the graph pair stimulus. The second column is a time stamp indicating when the RFV recorded the observation. The third and fourth columns indicate the x- and y-coordinates of the mouse. The rightmost column indicates if the content of the viewing window was visible (F) or obscured (B), or whether the mouse was not positioned over the graph pair at all (blank).

The decision participants made about whether the groups differed or not ("YES", "NO", "NO, they are about the same") is indicated in the row labeled "Response". The row "Time" showed the total time, in milliseconds, that participants spent viewing the stimulus before giving a response.

Procedure

Participants first read a brief description about the experiment and signed the consent form. They saw one practice graph pair with a brief explanation of what the circles in the graph represented and how to decode them (Figure 15). Participants demonstrated that they knew how to read the data correctly by interpreting data values in
the graph with respect to both attributes (e.g., "The four circles represent four people who have lived in Amherst for 1 year"). Only one participant had difficulty interpreting what the circles in the graph represented. In this case, I provided further instruction until she correctly interpreted the data.

Participants practiced moving the RFV viewing window around the practice graph using the mouse. When participants felt ready, they clicked on the computer screen to begin the experiment. Participants carried out the “visual search task” in the first 12 trials. The 12 graph pairs were presented in a random order to each participant. For each trial, participants first read a short description of the context for the data then clicked on “Next” to receive the question they had to answer (Figure 16a). After reading the question and clicking on "Next" again (Figure 16b), they examined the graph pair using the RFV (Figure 16c). Participants took as much time as they needed to view each graph pair. When they were ready, they clicked on one of three possible answers at the bottom of the screen. After giving a response, a screen appeared for 1500ms confirming their response, after which the next trial began.

After the visual search task, participants carried out the justification task. In these 12 trials, participants saw the 12 graph pairs again in the same order as in the visual search task. This time, however, they saw the entire graph pair on the screen unobscured by the RFV. The following instructions were presented on-screen:

You will see the graphs you saw earlier but this time the circles do not disappear. If you think you need to change your answer from what you gave earlier, you can change your answer here. Please read the questions and explain to the experimenter how you got your answers. Use the mouse to help show how you got your answers.

Participants took as much time as they needed to make their response.
In this graph, each circle stands for one person. The circle above the number '2' on the bottom line means that this person said he has lived in Amherst for 2 years. The three circles above the number '1.5' on the bottom line means that three people said they have lived in Amherst for 1.5 years.

Please tell the experimenter what the four circles above the '1' on the bottom line tell you.

Please tell the experimenter what the two circles above the '2' on the top line tell you.

Figure 15. Practice graph pair of hypothetical data. The graph shows the number of years that 21 individuals have lived in two different towns. Below the graph are a paragraph instructing participants how to interpret the graph and two questions to verify that participants could correctly interpret the meaning of circles in the graph.
The graph you will see shows data about the weights of fish from two species, the unmod and the genmod. Weight was measured in pounds. Each circle in the graph represents one fish.

a.

Do the unmods tend to be heavier than the genmods?

Please click on the NEXT button

b.

Figure 16. Three screenshots showing examples of what participants saw during each trial.

Continued, next page
Figure 16, continued

Do the unmods tend to be heavier than the genmods?

Please click on the button to give your answer

YES  NO  NO, they are about the same

c.

Interviewing Participants

The intention of the interview was to document the type of comparison methods participants offered to explain their decisions about the two groups. At each trial, participants read the question about the graph pair and were asked to “Please show the experimenter how you got your answer, then click your answer.” The interviewer initially remained silent to allow participants time to formulate a response. If after about 10 seconds a participant had not yet begun to speak, the interviewer used a prompt such as “Can you tell me how you decided whether one was shorter than the other?”
The experimenter also used prompts when a participant verbally reported a
decision about the graph pair without also specifying a comparison method. For example,
one participant said, "So now, do the trees given Q23 tend to grow more sprouts than
those given O27? I would say, ... the answer is no." Since the participant had not offered
a comparison method, the interviewer probed, "Because?" If participants made an
ambiguous reference to a graph location, (e.g., "because more circles are located towards
the lower length part of the graph," the interviewer would prompt so as to elicit a more
specific reference (e.g., "So the lower length, what do you mean lower length?" When
participants gave a reason or a phrase whose meaning was unclear, the interviewer often
repeated the unclear phrase, and asked them to clarify what they meant. For example, one
participant said that one group was greater than the other...

P: Because of the left shift.
I: The left shift?
P: Yeah.
I: What do you mean by that?
P: So...Q23 tend to grow more sprouts...right. Most of Q23 is less than most of
O27 so if you pick the...the median points or the bump points...

After the justification task, participants completed a short questionnaire
(Appendix B), and were then debriefed.

---

6 Expert participant PE014
7 Novice participant PN032
8 Expert participant PE014
Construction of Measures

Measure of Visual Search

RFV data from each trial can be converted into a "searchpath," a visual trace of where participants moved the viewing window during the visual search task. Figures 17a and 17b are examples of searchpaths from two participants. Each circle is an observation recorded by the RFV. The lines connect the dots in chronological order to create a trace of the path of the viewing window. Consecutive circles that are close together indicate slow mouse movements; ones that are far apart indicate fast mouse movements.

Horizontal Movement. To characterize how participants searched the stimuli, a measure of overall Horizontal Movement (HM) was calculated. The x and y coordinates of successive positions of the viewing window were recorded by the RFV during the visual search task. The distance traveled between any two consecutively sampled points can be described in terms of the changes in horizontal and/or vertical components of the mouse. From these individual measurements, I computed HM, the ratio of the total horizontal distance covered in a trial (dH) divided by the total distance traveled horizontally (dH) and vertically (dV):

\[ HM = \frac{dH}{dH + dV} \]
Figure 17. A searchpath from the novice participant 027 who, on the fish stimulus, used mostly horizontal movements (HM = 0.707). Each circle indicates the location of the center of the viewing window on the area of the graph pair at each moment recorded by the RFV. The circles are connected with blue lines to help better see the searchpath. The red arrow indicates where the participant began the visual search. The axes values show pixel locations.

Figure 18. A searchpath from the novice participant 105 who, with the soap stimulus, used mostly vertical movements (HM = 0.205).
Relatively large HM values indicate that participants searched a particular stimulus using more horizontal movements (i.e., “sweeping” movements). Relatively small HM values indicate searches with more vertical movements (i.e., “slicing” movements).

Values of HM are shown along with two searchpaths in Figures 17 and 18. Figure 17 is from a participant who used mostly horizontal sweeping movements, with a relatively high HM value of 0.707. Figure 18 shows a participant using relatively more vertical slicing movements. Her HM, 0.205, was relatively low.

**Measure of Right or Left Viewing Bias.** To test whether question wording affected where participants looked in a graph pair, I created a measure of viewing bias: the proportion of time a participant spent viewing the right side of a graph. To derive this proportion “PropTimeR,” I divided each graph pair into four quadrants (i.e., upper-left, upper-right, lower-left, lower-right), as shown in Figure 19. The top and bottom quadrants were divided at the horizontal axis of the top stacked-dotplot. The division between the corresponding left and right quadrants was placed at the median of each group. PropTimeR is the ratio of the time participants spent examining the right hand side of the graph pair divided by the total time participants spent examining the graph pair.

\[
\text{PropTimeR} = \frac{\text{time in top right quadrant} + \text{time in bottom right quadrant}}{\text{Response Time}}
\]

Figure 20 shows the searchpath for a novice participant who spent most of the time viewing the left half of the graph pair, resulting in a low PropTimeR value of 0.005.

---

\[9 \text{ Response Time (RT) was the total time a participant spent on a trial before making a response.}\]
Figure 19. The four quadrants of the "Fish" stimulus

Figure 20. An example of a participant with low PropTimeR (.005). This searchpath belongs to novice participant number 24. The red arrow indicates where the participant began her visual search. In examining the stimulus “towntemps,” she spent almost all her time on the left of the graph pair. The question used the “less” phrasing, which I predicted would induce some novices to spend more time viewing the left of the graph pair. In the justification task, this participant also made verbal references to data located in the left hand side of the graph pair.
Recall that the three conditions manipulated in the visual search and justification tasks were: (1) expertise – novice, expert; (2) question wording – more, less; (3) meandiff – small, medium, large. Although the justification task was given to participants after the visual search task, I present first the results of the justification task.

**Justification Task**

Development of a Coding Scheme for the Justification Task

The purpose of the justification task was to gain evidence about the comparison method participants used to decide about groups differences. In preparation for coding the interviews, complete transcripts were made of participants’ verbal responses during the justification task. Four of the categories used for coding the transcripts were based on prior research. These included the categories global displacement, global cutpoint, local cutpoint, and local slice. Four other coding categories were developed for use in this study: global other, local other, local extremes, and other.

The three "global" categories included responses that used all the data cases, for example, discussing data proportionally or using aggregates. The four "local" categories included responses that used sub-groups of data, for example, comparing absolute numbers of case with respect to cutpoints or within slices. The "Other" category was used to categorize responses that could not be coded as either global or local, or were a mixture of global and local responses. Table 2 includes details on the coding scheme along with examples.
<table>
<thead>
<tr>
<th>Sub-Category</th>
<th>Criteria</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global Displacement</td>
<td>The participant referred to the presence or absence of a displacement between the groups using location of aggregate features such averages (mean, median, mode), modal clumps, or for general shifts in the entire distribution as indicated by the range (complete range or truncated range) of the data.</td>
<td>&quot;...because of a modest shift in the gentle bump for cold [days] is sort of shifted a little bit to the left of the gentle bump for the number of hot days.&quot; (Expert 14, roadkill graph pair)</td>
</tr>
<tr>
<td>Global Cutpoint</td>
<td>The participant placed a cutpoint across both groups and compared the proportion/percentages of the data in one group to the proportion/percentage of data for the other group. Participants may have referred to data on just one or both sides of the cutpoint.</td>
<td>&quot;...the highest point of the bell curve was around looks like about 12 or 13 minutes, and it was for seven, six patients. And uh, but the highest point of the bell curve was only three minutes for Zolcain. So Zolcain took shorter time, so, yes, Pandol took longer time&quot; (Novice 16, numbness graph pair)</td>
</tr>
</tbody>
</table>

Table 2. Coding scheme used for coding participant justifications.
<table>
<thead>
<tr>
<th>Table 2, continued</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Global Other</strong></td>
</tr>
<tr>
<td>The participant used a global explanation that did not fit the above global categories.</td>
</tr>
<tr>
<td>“...because most of their [Pandol] patients... took anywhere from 7½ minutes all the way up to 20 minutes to numb versus the Zolcain patients who are lower on the scale, and only experienced up to 15 minutes... before numbness begins” (Novice 4, numbness graph pair)</td>
</tr>
<tr>
<td><strong>Local Slices</strong></td>
</tr>
<tr>
<td>The participant isolated one or several contiguous stacks of data in the two distribution, using the same lower and upper x-axis boundaries across both distributions. They then compared the frequency of cases of the two groups in those stacks, ignoring the rest of the data. The slice may have included stacks/cases at the endpoints and there was no restriction for how large a slice could be.</td>
</tr>
<tr>
<td>“Q23 grow fewer sprouts than the trees with O27 because the number of [O27] sprouts are bigger, I would say, from 5 to 10 are more than Q23 is from 5 to 10” (Novice 24, treesprouts graph pair)</td>
</tr>
<tr>
<td><strong>Local Cutpoint</strong></td>
</tr>
<tr>
<td>The participant placed a cutpoint across both groups and compared the number of cases in each group with respect to the cutpoint.</td>
</tr>
<tr>
<td>“After about 10 there is only 4 circles, 4 wires in aluminum whereas for copper there is many after 10 so I said that copper wires tend to twist more than aluminum wires.” (Novice 25, modulus graph pair)</td>
</tr>
<tr>
<td><strong>Local Extremes</strong></td>
</tr>
<tr>
<td>The participant compared the number of cases at one extreme location (i.e., minima, maxima, left- or right-tails of the distribution) of one group to the number of cases at that same location in the other group.</td>
</tr>
<tr>
<td>“Because it takes at least 7 minutes for a patient to start and for most patients it takes even longer than 10 minutes where for Zolcain it starts right away sometimes...” (Novice 104, numbness graph pair)</td>
</tr>
</tbody>
</table>

Continued, next page
<table>
<thead>
<tr>
<th>Local</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>The participant used a local explanation that did not fit any of the above local categories.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>• The participant used an explanation that did not fit any of the above categories;</td>
</tr>
<tr>
<td>• Or the participant mixed global and local comparison methods in one explanation, for example, s/he compared the mode of the first group with the minima of the second group;</td>
</tr>
<tr>
<td>• Or the participant did not give an answer or an explanation at all.</td>
</tr>
</tbody>
</table>

| “it still has this one dot here at 22 and there is no dots on 22 for Sabun...[participant-experimenter dialogue omitted]... At 13 they are the same and at 14 they are the same and then Fab soap has a little bit less at 15 and they are tied again but then it catches up in this category. It is kind of tricky but I would say that... I guess Sabun has more...” (Novice 29, soap graph pair) |

| “I kind of just saw the shape of the graph and I kind of...I based my answer only on that. How can I explain this? Like only on the shape, I wasn’t actually... I kind of like, I just compared the two graphs. I... I don’t have words for it” (Novice 111, roadkill graph pair) |

| • mixing range with minima “copper tends to twist more just because it is further down. It starts at 5 and goes to like 17 and this one starts lower” Novice 24, modulus graph pair) |

| • “I am just really guessing. I am looking at the graph and what I am seeing from the graph, I am guessing. I am making...I am guessing an answer.” (Novice 103, frozen graph pair) |
Coding Transcripts

To code a transcript, the coder read from the beginning of a transcript until the participant gave a definitive answer as to whether the groups differed. Then the coder located the first comparison method that the participant used to justify the answer (the justification). The justification could appear either before or after the participant’s answer. If one comparison method appeared immediately before the answer, and another appeared immediately after, the coder would code the justification that came before the answer.

The coder also noted whether the coded justification made reference to either of the four quadrants (top-left, top-right, bottom-left, bottom-right quadrants) in the graph pair. When the justification referred to data on the left only, it was coded as “LeftTailed.” If it referred to data on the right only, it was coded as “RightTailed.” If a justification made no references to the left or right, or referred to both the left and right, then the justification was coded as “No Tails”. Two examples of how transcripts were coded are shown in Appendix C.

Coder 1 coded all the transcripts. To estimate the reliability of coding, a second coder (Coder 2) independently coded a randomly selected sub-set of the transcripts (16% of them). Both coded the transcripts blindly with respect to whether the participant was an expert or a novice. However, Coder 1 coded all 12 trials from a single subject together. This may have produced a bias in the coding in the sense that earlier trials from a participant may have subtly influenced the coding of later trials. Coder 2 did not receive the transcripts organized by participant, so did not have such knowledge.
Justification Task Results

Although participants were not explicitly asked to make judgments about the statistical significance of the difference between the two groups, their verbal responses indicated that they responded as expected: the greater the meandiff, the greater the number of responses stating that groups differed.

Verbal Response Correctness

A “correct” response in the large and medium meandiff conditions was a response that matched how the group means were positioned relative to one another in the graph pair. In the small meandiff condition, every response was considered correct\(^{10}\). Using these criteria, both experts (mean correct out of 12 responses = 11.5, 95%CI: 11.1, 11.9) and novices (mean = 11.4, 95%CI: 11.1, 11.6) performed at very near the maximum level. This suggests that participants, including the novices, were attending well to the task.

Use of Local and Global Comparison Methods

The distribution of novice and expert comparison methods are shown in Figure 21, based on the codings of Coder 1. I predicted that experts would tend to use more global comparison methods than novices. The results confirmed this prediction. Almost all expert responses were global. Novice responses were distributed about equally among the global, local, and other categories. However, the overall interrater agreement between Coder 1 and 2 in coding justification types was fairly low\(^{11}\) at \(\kappa = 0.54\)\(^{12}\) (ASE=0.07).

---

\(^{10}\) All responses were acceptable because participants could decide either that they saw an absolute difference between the two group means or that the two means were not significantly different from one another and so were the same.

\(^{11}\) 0.40 and 0.79 indicate fair to good, SYSTAT v.11 Help Manual.
A factor contributing to this low value of interrater reliability was the difficulty in identifying the same excerpts of the transcript to code. Some responses were very short, and in these cases both coders identified the same part of the transcript to code as the justification. Other responses were longer, and sometimes included multiple comparison methods. This made it more difficult to isolate the same piece of the transcript to code as the justification. (See Appendix C, Example 2 for a response with multiple comparison methods that was difficult to code). Long responses, combined with multiple comparisons methods used in a single transcript, reduced the chance that both coders identified the same portion of the transcript to code.

![Figure 21. The distribution of comparison methods for expert and novice participants.](image)

12 This statistic may be suspect because more than one-fifth of the cells in this table contain counts that are less than five.
Figure 22. The distribution of comparison methods for novice participants, as coded independently by two coders.

Figure 23. The distribution of comparison methods for expert participants, as coded independently by two coders.
Despite the low value of interrater reliability, both coders showed general agreement in terms of the distribution of comparison methods used by novices (Figure 21) and experts (Figure 22). Both coders found that experts and novices used Global Displacements the most, while the Local Slice was the most prominent novice local comparison method.

**Effect of Question Wording on Comparison Method Used**

No prediction was made about how question wording (i.e., “more” or “less”) might influence the comparison method used. The results give no evidence $\chi^2 (7) = 3.63$, $p = 0.82$) of there being an effect for either the expert ($\chi^2 (2) = 3.07$, $p = 0.22$) or novice, $\chi^2 (7) = 2.74$, $p = 0.91$)

**Effect of Question Wording on Direction of Tail References**

I predicted that novices (but not experts) would be influenced by the “more” or “less” question wording, incorporating into their verbal responses more references to subsections of data in the left or right quadrants of the graph pair. Of the 480 total participant responses, 79 included a tail reference. All were from novices (Table 3). The majority of these (59) mentioned the right tail.

Justifications were coded as to whether they made no exclusive references to data in either tails of the graph pair, included references only to data in the left, or included references only to data in the right. The interrater agreement in categorizing of tails was low at $\kappa = 0.63$ (ASE = 0.108). This probably was for the same reasons that the coding of justifications into comparison methods above was low in reliability – the coders had difficulty identifying the same parts of the justification to code.
Table 3 shows that, regardless of question wording, there was a tendency for novices to refer more to the right than to the left tail ($\chi^2 (1) = 19.3, p = 0.000$). However, the overall $\chi^2$ performed on the 2x3 table in Table 4 suggests that question wording had no effect on whether a particular tail was mentioned or not in the justification task, ($\chi^2 (2) = 3.65, p = 0.161$).

<table>
<thead>
<tr>
<th>Question wording</th>
<th>Verbal Response</th>
<th>NoTail (%)</th>
<th>LeftTail (%)</th>
<th>RightTail (%)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>less</td>
<td></td>
<td>148 (81.8)</td>
<td>10 (5.5)</td>
<td>23 (12.7)</td>
<td>181</td>
</tr>
<tr>
<td>more</td>
<td></td>
<td>133 (74.3)</td>
<td>10 (5.6)</td>
<td>36 (20.1)</td>
<td>179</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>281 (78.1)</td>
<td>20 (5.6)</td>
<td>59 (16.4)</td>
<td>360</td>
</tr>
</tbody>
</table>

Table 3. Table showing the number of novices who mentioned no tails, the LeftTail, or the RightTail in response to a “less” or “more” worded question. The percentages are row percentages.

Table 4 shows the number of tail references for each stimulus and for each wording of the question. Responses to the numbness stimuli are particularly striking. The numbness stimulus asked participants “Does Pandol tend to take a longer/shorter time to numb a patient than Zolcain?” With the more phrasing of the numbness question, 7 participants gave a justification that mentioned the left tails of the graphs; none mentioned the right tails, as we might have expected. This unusual pattern of responses may have been due to two features of the numbness problem. First, unlike most of the stimuli, the “desirable” values in the numbness context are on the lower end of the scale (shorter times to become numb). This low-end preference may lead to a tendency, regardless of problem wording, to look at the lower end of the scale. But secondly, the axis of the graph pair was misleadingly labeled as “Time for numbness to begin.” It should instead have been labeled either “Time for drug to take effect” or “Time for
numbness to be complete." This mislabeling may have further increased participants’
tendencies to look toward the lower end of the scale (to times when numbness “began”).

<table>
<thead>
<tr>
<th>GRAPH</th>
<th>Verbal Response</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>less</td>
</tr>
<tr>
<td></td>
<td>LeftTail</td>
</tr>
<tr>
<td>fish</td>
<td>3</td>
</tr>
<tr>
<td>frozen</td>
<td>1</td>
</tr>
<tr>
<td>modulus</td>
<td>1</td>
</tr>
<tr>
<td>numbness</td>
<td>1</td>
</tr>
<tr>
<td>petals</td>
<td>1</td>
</tr>
<tr>
<td>recharge</td>
<td>1</td>
</tr>
<tr>
<td>roadkill</td>
<td>2</td>
</tr>
<tr>
<td>Soap</td>
<td>6</td>
</tr>
<tr>
<td>spintops</td>
<td>2</td>
</tr>
<tr>
<td>termites</td>
<td>2</td>
</tr>
<tr>
<td>towntemps</td>
<td>2</td>
</tr>
<tr>
<td>treesprouts</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 4. The number of LeftTail and RightTail responses mentioned in the 12 different graph pairs given more- or less-worded problems.

For the above reasons, I repeated the analyses after removing the numbness problem. These results are summarized in Table 5. Based on this reanalysis, it seems that question wording did have an effect of verbal response, \( \chi^2 (2) = 7.17, p = 0.03 \) Novices were about three times more likely to use a LeftTail reference with the “Less” question compared to the “More” question. They were about twice as likely to use a RightTail reference with the “More” wording.

<table>
<thead>
<tr>
<th>Question wording</th>
<th>Verbal Response</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NoTail (%)</td>
</tr>
<tr>
<td>less</td>
<td>138 (81.7)</td>
</tr>
<tr>
<td>more</td>
<td>122 (75.8)</td>
</tr>
<tr>
<td>Total</td>
<td>260 (78.8)</td>
</tr>
</tbody>
</table>

Table 5. Table 3, with the responses to the Numbness context removed, showing the number of novices who mentioned no tails, the LeftTail, or the RightTail in response to a “less” or “more” worded question.
This result lends some support to the possibility that some participants might have been misinterpreting questions that use a "more" or "less" phrasing as asking only about data in the right or left tails of the distribution. For example, novice participant 24 was shown the graph in Figure 22a and asked if more animals were killed on cold days than on hot days. She answered,

P: My first impression is that there are more roadkill on cold days because there are more at the beginning but the end is really what you need to be looking at.
I: Should be looking at the end?
P: Yeah because they are the highest numbers.

This participant argued that only cases on the right of the graph should be examined because that was where the "highest numbers" were, presumably meaning the higher number of animals killed in a day.

PN027 was shown the graph in Figure 22b and asked whether Top A tended to spin for a longer time than Top B. Asked to justify her answer, she responded:

P: I just tried to compare the two and see if there were more circles in a closer to the end of, you know, for longer times to see if there were more dots in a longer time period...
I: You were looking at all of the circles or the circles at the end or?
P: Well I went over all of them. I think I went over all of them but I think the important ones were really just at the end, I think.

Again, PN027 expressed a sense that the data cases in the right quadrants, corresponding to the tops that spun longer, were the important ones to consider.
Taken together, these excerpts support Konold, Pollatsek, Well, and Gagnon’s (1997) speculation that some participants appeared to be reinterpreting a question about the global tendencies of two groups as a question about only a subset of the group (e.g., the lighter fish or the heavier fish). The current study further supports this possibility by showing that the wording of the question can influence which subset of the data participants focus on.

Effect of Meandiff Level on Comparison Methods

No predictions were made about how meandiff might affect the type of comparison method used in the justification task. Table 6 shows the percentage of global, local and other comparison methods used at each meandiff condition (collapsing over the various subcategories).

A $\chi^2$ test found a significant association overall between meandiff and comparison methods ($\chi^2 (4) = 33, p < 0.000$). Novices ($\chi^2 (4) = 47.0, p < 0.000$), but not experts ($\chi^2 (2) = 3.28, p = 0.194$), appeared to use different comparison methods depending on the degree of overlap in the two distributions. For stimuli where the difference between the
means of the two groups was large, novices tended to use global comparison methods, with about twice the frequency they used when the difference between the means was small or medium. This finding adds support to a speculation by Konold and Pollatsek (2002) that it might be easier for novices to perceive averages as group properties, and thus to use them to compare two groups, when the two distributions have little overlap.

<table>
<thead>
<tr>
<th>Meandiff Method</th>
<th>Novices</th>
<th></th>
<th></th>
<th>Experts</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Small</td>
<td>Medium</td>
<td>Large</td>
<td>Small</td>
<td>Medium</td>
<td>Large</td>
</tr>
<tr>
<td>Global</td>
<td>35.8</td>
<td>28.3</td>
<td>66.7</td>
<td>90.0</td>
<td>95.0</td>
<td>82.5</td>
</tr>
<tr>
<td>Local</td>
<td>34.1</td>
<td>49.2</td>
<td>15.8</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Other</td>
<td>30.0</td>
<td>22.0</td>
<td>17.5</td>
<td>10.0</td>
<td>5.0</td>
<td>17.5</td>
</tr>
</tbody>
</table>

Table 6. Percentages for types of comparison methods used in the justification task by expertise meandiff conditions. Each column sums to 100%.

Visual Search Task

Data from four participants, all novices, were corrupted during the processing of their files. Trials from these four participants are thus not included in the analyses of the visual search task.

Visual Search Task Results

Response Time Effects

Response Time (RT) on each trial was measured from the moment data were first visible in the RFV viewing window until the time a participant clicked on a response on the computer screen. Past studies (e.g., Reingold, Charness, Pomplun, & Stampe, 2001) have shown an expert advantage in response times when people had to make same-different judgments for two meaningful visual stimuli (e.g., chess pieces on a
chessboard). In the current study, no significant expert advantage on RT was found. Experts and novices spent roughly equal amounts of time making their decisions about the two groups \( t(22) = -0.721, p = 0.478 \), with experts taking on average 27.9s (95%CI: 20.7, 35.0) and novices 30.9s (95%CI: 25.4, 36.3). The lack of an expert-novice difference may have been partly due to the demands of using the RFV viewing window, which may have served to put experts and novices on somewhat equal footing with regard to performing the rudimentary aspects of the task.

However, there was a slight practice effect for novices, but not for experts. The mean RT for the novices’ first 6 trials was 35.6s (95%CI: 29.4, 41.7), compared to 26.1s (95%CI: 20.9, 31.4) for the last 6 trials, \( t(25) = 5.45, p = 0.000 \). Experts’ first 6 trials had a mean RT of 29.4s (95%CI: 22.8, 36.1), compared to 26.3s (95%CI: 17.5, 35.2) for the last 6 trials, \( t(9) = 1.12, p = 0.29 \). The decreasing RT for novices may indicate that over trials, novices were becoming more familiar and efficient at the visual search task. Alternatively, it could be that in early trials, novices where formulating a strategy for how to compare two groups, while experts already had a strategy, which they used consistently throughout the trials.

An ANOVA found no RT difference based on question wording across experts and novices, \( F(1, 34)=0.49, p=0.49 \), but a significant main effect of meandiff on RT \( F(2, 68) = 17.1, p = 0.000 \). As shown in Table 7, RT increased as meandiff became smaller, and this general pattern held for both experts and novices. Contrasts showed that the large meandiff condition was responded to significantly faster than the medium meandiff condition, \( F(1, 34) = 7.80, p=0.009 \), which in turn was responded to significantly faster than the small meandiff condition, \( F(1, 34) = 11.9, p=0.002 \). Thus, for both experts and
novices, questions about distributions with a high degree of overlap took more time to answer than did questions about distributions where the means were further apart.

<table>
<thead>
<tr>
<th>Meandiff</th>
<th>Overall</th>
<th>Experts</th>
<th>Novices</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>95% Confidence Interval</td>
<td>Mean</td>
</tr>
<tr>
<td>Large</td>
<td>24.0</td>
<td>20.4, 27.6</td>
<td>18.7</td>
</tr>
<tr>
<td>Medium</td>
<td>29.0</td>
<td>24.7, 33.2</td>
<td>27.1</td>
</tr>
<tr>
<td>Small</td>
<td>37.1</td>
<td>30.2, 44.0</td>
<td>37.8</td>
</tr>
</tbody>
</table>

Table 7. Mean and confidence intervals for RT in the three meandiff conditions.

The only significant interaction was between meandiff and wording (F(2, 68) = 4.00, p = 0.023). Contrasts showed that this interaction was mostly due to the interaction between wordings in the small versus medium meandiff conditions, F(1, 34) = 7.52, p = 0.010. Response time in the small meandiff condition increased from 33.6s (95%CI: 29.6, 39.8) in the less wording to 39.2s (95%CI: 32.1, 46.3) in the more wording. By contrast, response time in the medium meandiff condition decreased from 31.9s (95%CI: 26.6, 37.2) in the less wording to 25.6s (95%CI: 21.9, 29.3) in the more wording.

The finding regarding response time and mean difference may not generalize to viewing graphs more generally. The different RTs over the different meandiff levels might be largely explained in terms of perceptual discrimination difficulties peculiar to looking at data using the restricted view window. With distributions that are far apart, the fact that there are differences of some sort usually becomes apparent in the first few positionings of the window. This is not true of distributions with high overlap.
Effects of Expertise, Question Wording, and Meandiff on Horizontal Movement

Experts, who were expected to compare groups using global comparison methods, were therefore expected to use more horizontal search movements while inspecting the graphs. This was based on the assumption that locating the average of each group would most efficiently be accomplished by moving the viewing window back and forth horizontally across each group separately. Novices, who were expected to use local comparison methods, were expected to use more vertical search patterns. This was based on the assumption that moving the viewing window vertically would be the most efficient way to locate and evaluate slices or cut points.

An ANOVA testing the HM effects of expertise, question wording, and meandiff supported these hypotheses. Average expert HM (0.65, 95%CI: 0.59, 0.72) was greater than the average novice HM (0.59, 95%CI: 0.57, 0.62), F(1, 34) = 5.57, p = 0.024. The average HM for all participants was 0.61 (95%CI: 0.59, 0.63). There was no HM practice effect for experts or novices.

Question wording showed no effect on HM, F(1, 38)=0.21, p = 0.647. The mean HM in the less-phrased condition (0.61, 95%CI: 0.60, 0.63) was similar to the mean HM in the more-phrased condition (0.60, 95%CI: 0.59, 0.62). Nor was there an interaction between expertise and question wording, F(1, 34) = 0.214, p = 0.647. This result is consistent with the results reported earlier concerning the lack of a question-wording effect on the type of comparison methods used. Regardless of how a question was worded, participants seemed to use the same general visual search pattern (See Appendix D).
The mean HMs across the three meandiff conditions were not significantly different \( F(2, 68) = 2.934, p = 0.060 \), though the fact that the pattern of the means holds up for both experts and novices is intriguing (Figure 25).

![Graph showing expert and novice HM and 95% confidence intervals for the three meandiff conditions.]

**Figure 25.** Novice and expert HM and 95% confidence intervals for the three meandiff conditions.

The pattern suggests that for distributions with means that were fairly similar, both novices and experts used somewhat more vertical searching than they did for distributions with means that were far apart. If there is an effect here, the explanation for
the effect would probably differ for experts and novices. For novices, it may be that as
the means of two distributions become farther apart, they are more likely to adopt an
expert-like perspective, examining the data as a whole using a sweeping horizontal search
pattern. This finding would be consistent with the earlier result showing that the larger
meandiff conditions elicited more global verbal responses in novices than did the smaller
meandiff conditions. For experts, it may be that as the means of two groups get very
close, they spend somewhat more time moving vertically between the two distributions,
perhaps trying to judge the relative displacement between the means of the two groups.

**Effect of Question Wording on Search Location**

As discussed previously, the results of the justification task indicated that question
wording influenced the probability of novices making exclusive reference to one side of
the graph stimuli in justifying their answer. To see whether question wording similarly
influenced visual search behavior, I examined the proportion of time participants spent in
the right versus left quadrants of the stimuli given the two phrasings of the question.

Overall, Mean PropTimeR for experts and novices was not significantly different:
0.545 (95%CI: 0.478, 0.613) for experts compared to 0.479 (95%CI: 0.443, 0.515) for
novices, t(15.6) = 1.92, p = 0.07. Table 8 shows PropTimeR for novices and experts
broken down by question wording (See Appendix E). Although experts spent more time
on the right than they did on the left regardless of wording, this difference was not
significant t(9) = -1.47, p = 0.18. The slightly higher than 50% time spent by experts
overall on the right could, of course, have been just due to chance. However, it could also
be due in part to an artifact of the way I chose to divide the graphs into the left and right
quadrants for this analysis using the median. Examining the stimuli further, I found that
the mean width of the right quadrant (1699 pixels) was about 15% greater than the mean width of the left quadrant (1469 pixels).

In contrast to the experts, novices spent significantly more time on the right with questions phrased in terms of “more” (0.516) than with questions phrased in terms of “less” (0.442), $t(25) = -4.62$, $p = 0.000$. This supports one of the major predictions of the study, and I discuss some implications of this finding later.

<table>
<thead>
<tr>
<th>Question wording</th>
<th>Expert</th>
<th>Novice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less</td>
<td>0.522 (0.427, 0.616)</td>
<td>0.442 (0.408, 0.476)</td>
</tr>
<tr>
<td>More</td>
<td>0.569 (0.516, 0.623)</td>
<td>0.516 (0.471, 0.560)</td>
</tr>
<tr>
<td>Mean difference</td>
<td>-0.048 (-0.121, 0.026)</td>
<td>-0.073 (-0.106, -0.041)</td>
</tr>
</tbody>
</table>

*Table 8.* Mean proportion of time viewing the right side of the graph pair (PropTimeR) and 95% confidence intervals given question wording and level of expertise.

### Two Types of Novices

I have been assuming to this point that all of the undergraduates were novices. This seems, however, an unreasonable assumption. Indeed, we saw from the analysis of the verbal justifications that many novices appeared to use global methods of justifying their answers. To explore whether there might be two types of undergraduates in the novice group — truly novice undergraduates and more expert-like undergraduates — I separated the undergraduates according to the number of tail references they made. This was based on the assumption that the reference to tails in justifying a group difference was the most salient difference between experts and novices: No expert ever made such a justification while many novices did. The 16 undergraduates who mentioned tails in two or more justification trials were classified as “novice undergraduates” while the 10
undergraduates who made 0 or 1 such justification were considered “expert-like undergraduates.”

The number of justification responses for which a tail-based justification could be mentioned ranged from 0 to 12. On average, expert-like undergraduates mentioned tails 0.80 times (95%CI: 0.50, 1.10) in the 12 justification trials compared to 3.81 times (95%CI: 2.82, 4.81) for the novice undergraduates. The average number of mathematics classes that novice undergraduates had taken was 1.8 (95%CI: 1.4, 2.1) versus 2.5 (95%CI: 1.7, 2.5) for expert-like undergraduates. This difference in means was not significant, t(19.5) = -1.520, p = 0.145. Both novice- and expert-like undergraduates had taken about 1.7 statistics classes.

I repeated two of the analyses using level of undergraduate expertise as a variable to examine how they differed, if at all, in the visual search tasks. First, I conducted a 2 (expertise: novice undergraduates vs. expert-like undergraduates) by 3 (meandiff: low, medium, high) by 2 (question wording: more vs. less) ANOVA using HM as the dependent variable. The mean HM for the novice undergraduates (0.56, 95%CI: 0.53, 0.60) was significantly lower than the expert-like undergraduates (0.64, 95%CI: 0.61, 0.66), F(1, 24) = 11.7, p = 0.002. There were no other main or interaction effects. This result suggests that some of the undergraduates were expert like, in that they both rarely referred to tails in their justifications and they tended to use visual searches that more closely resembled those of the expert (recall that the mean HM for the faculty experts was 0.65).

I repeated the ANOVA above using PropTimeR as the dependent variable. Only the question wording variable showed a significant effect, with participants overall
spending more time examining the right-hand side of the graph, F(1, 24) = 7.41, p = 0.012 (see Table 8 for the overall undergraduate means). The expert-like undergraduates were not less influenced by the question wording than were the novice undergraduates.

Based on the analyses of the HM and PropTimeR together, the evidence is mixed on the question of there being two types of undergraduates.

A Possible Gender Confound

In this study, the novice participants were mostly female (26 of 30) while the expert participants were mostly male (8 of 10). A possible gender confound cannot be ruled out as a factor to explain the different visual search and justification results. Although a further study controlling gender would need to be carried out to conclusively discount this possibility, I present some evidence to suggest that the observed differences were not due to gender.

Table 9. Mean HM and PropTimeR (and their 95% confidence intervals) of experts and novices. Note that four novice participants have been excluded from this table.

<table>
<thead>
<tr>
<th>Expertise Level</th>
<th>Dependent Variable</th>
<th>HM</th>
<th>PropTimeR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female (N=2)</td>
<td>0.673</td>
<td>(-0.132, 1.477)</td>
<td>0.552</td>
</tr>
<tr>
<td>Male (N=8)</td>
<td>0.648</td>
<td>(0.568, 0.727)</td>
<td>0.543</td>
</tr>
<tr>
<td>Novice</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female (N=23)</td>
<td>0.586</td>
<td>(0.558, 0.614)</td>
<td>0.480</td>
</tr>
<tr>
<td>Male (N=3)</td>
<td>0.630</td>
<td>(0.590, 0.669)</td>
<td>0.474</td>
</tr>
</tbody>
</table>

Table 10 presents the mean HM and PropTimeR of male and female participants according to their expertise (after removing the 3 female and 1 male novice with incomplete data). Both expert and novice females had HM values that were fairly similar
to their male counterparts (experts: $t(8) = 0.34, p = 0.747$, novices: $t(24) = -1.133, 0.268$).

They also had similar values in PropTimeR (experts: $t(8) = 0.115, p = 0.911$; novices: $t(24) = 0.109, p = 0.914$. Because of the small numbers of values in these comparisons, they are highly unstable. But together they offer some evidence that the difference between experts and novices is not due primarily to gender.
CHAPTER 4

GENERAL DISCUSSION

Most of the research to date exploring the methods that novices use to compare groups has dealt with the responses participants offer when comparing groups (e.g., Gal, et al., 1989; Watson & Moritz, 1999; Konold, et al. 2002). While this research has indicated that novices use non-normative “local” methods to compare groups, one could argue that what novices offer as an explanation may not reflect their underlying way of thinking about the data. Perhaps novices arrive at their decisions about group differences just as experts do – for example, by visually judging the relative location of the centers of the distribution. But when asked to verbally justify or explain their decisions, novices may resort to more simplistic and invalid rationales – for example, comparing the number of cases in a slice, which can easily be used to make a numerical comparison between the two groups.

This study investigated expert and novice data analyzers as they compared two groups. Furthermore, it explored both how they visually searched graphed stimuli and how they subsequently justified their decisions about whether the groups differed. Finding an association between data analyzers' visual search behavior and their verbal justifications would provide evidence that both of these measures reflect an underlying perspective about the nature of the task of group comparison. Specifically, people who take a statistical perspective regarding the task would consider the collection of cases as an aggregate, and thus look for features that apply to the aggregate, such as the mean or shape of the distribution. One would therefore expect that they would structure their visual inspection of a graph in a way that would allow them to easily identify these
aggregate features. Furthermore, we would expect that their verbal justifications for why the groups differed or not would highlight global features of the data. Those not taking a statistical perspective would, on the other hand, attend to more local features of the data, such as the location of extreme values, and this local approach to the data would be evident in both their verbal justifications and in how they visually inspected the graphs.

With regard to how experts versus novices would visually search the graph pairs, I predicted that experts would use more horizontal movements, structuring their visual searches to find the aggregated features of each group. For novices, I predicted they would use more vertical movements as they searched for cutpoints and slices. I also expected that the visual search patterns of novices (but not of experts) would be affected by the wording of the question, such that the time novices spent looking at one side of the stimulus would be influenced by whether the question was worded in terms of which group was more or which group was less.

With regard to the verbal justifications, I predicted, that novices would use mostly local comparison methods while experts would use mostly global comparison methods. I furthermore hypothesized that some novices (but not experts) would be influenced by the words "more" and "less" in the wording of the question and would use verbal justifications that included reference to the corresponding tails of the graph pairs.

All these predictions were verified.

Limitations of the Study

There were a number of limitations to this study that future investigations should address. Recall that participants were asked to carry out the visual search task for all of
the stimuli before proceeding to the verbal justification task. This sequential approach was used to avoid the possibility that participants would adjust their visual search pattern in response to the verbal justification task. However, a drawback to this approach is that by the time participants got to the verbal justification task, they sometimes had forgotten how they had answered the question in the visualization task. Novices, in particular, may have thus approached the problem in the justification task differently than they did in the visualization task. Furthermore, during the visualization task there was some evidence that novices were evolving a strategy during the earlier trials, as suggested by the fact that as the trials progressed, they spent less time, on average, on a trial. Thus the responses to the justification tasks may have been somewhat different had each trial involved first a visual scan followed immediately by the justification task. A study that included both this "integrated" presentation, as well as the sequential design used in this study, would help establish whether this ordering is important or not.

In this study, the number of cases in each group of a graph pair were equal. This reduced the number of factors manipulated. It also probably made the task easier for many of the novices. Gal, Rothschild, and Wagner (1989) gave sixth graders graph pairs in which the two groups had similar means but different numbers of cases. In this instance, less than half of the students were able to correctly judge group differences; they had little difficulty when the numbers of cases in the two groups were equal.

One drawback of using equal-sized groups was that it was harder to tell in some instances whether or not participants were using proportional reasoning when comparing the groups. This was especially difficult with the cutpoint strategies, which were considered global when used to reason about percentages of cases but local when used to
compare absolute number of cases. To be able to assess more accurately whether participants are using local or global reasoning, future studies might use stimuli that include examples where the number of cases in the two groups are noticeably different.

As mentioned previously, the unfortunate confounding of gender with expertise leaves open the possibility that the differences attributed in this study to level of expertise were due in part (or even entirely) to gender differences. While this seems implausible, it cannot be ruled out.

Finally, the low inter-rater reliability in coding the verbal responses into justification types calls into question the conclusions about difference between how experts and novices justify their conclusion. As mentioned previously, one of the major reasons for disagreements between the two coders was that they frequently identified different sections of the transcript to code as justifications. And because novices in particular often used several different justifications, this led to differences in categorizing their responses. One possible solution would be to structure the interviews more carefully to make it easier to identify a single justification to code, perhaps by asking participants to justify their answer before proceeding to ask them additional questions.

Conclusions, Implications, and Future Research

Despite the limitations of the study, the results showed that, as predicted, experts used relatively more horizontal sweeping search movements than novices in the visual search task, and furthermore were more likely than novices to describe global comparison methods in justifying whether the groups differed.
The prediction that question wording would influence how novices verbally responded to the group comparison task was somewhat confirmed, although the results should be interpreted with caution. While the proportion of verbal responses that made specific references to sub-sections of the graph pair was small, almost every novice included at least one such reference. There was a tendency for novices to make references to data on the right hand side of the graph. Also, after removing the data from one of the problems that appeared highly unusual, the relative proportion of time novices made references to the left did increase with the “less” wording and decreased with the “more” wording.

Question wording also affected where novices searched the graph pair. Novices spent more time examining the right-hand side of the graph pair when the “more” wording was used, and spent more time on the left-hand side when the “less” wording was used. This suggests that some novices misinterpreted questions about group properties to be questions about a sub-set of the data. For example the question “Do the unmod fish tend to be heavier than the genmod fish?” may have been interpreted by some novices as the question “Among the heavier fish, are there more unmods or genmod fish?”

The correspondence between how novices responded in the justification and visual search tasks provides more evidence that both verbal response and visual search behavior can be used for assessing how people reason about data. The fact that question wording had an effect on how novices, but not experts, approached data in the two tasks argues against the possibility mentioned previously—that novices approach the group
comparison task in the same way as experts but then offer simplistic verbal explanations that do not reflect how they actually accomplished the task.

An unanticipated finding was that search behavior for both experts and novices was affected by the degree of separation between the means of the two groups shown in a graph pair (i.e., the “meandiff”). Data analyzers used relatively more horizontal search movements when examining the large and medium meandiff stimuli than they did with the small meandiff stimuli. Furthermore, novices increased their use of global comparison methods in the large meandiff condition compared to the other two conditions. It seems that when the two groups were farther apart, novices were more likely to treat each distribution as a group, describing and comparing them using aggregates. As was mentioned, this finding supports a speculation offered by Konold and Pollatsek (2002) that it might be easier for novices to perceive and use aggregate features of distributions in the two-group comparison task when the distributions are relatively far apart.

This study may have something to contribute to a more general theory of graph comprehension. Trickett, Ratwani, and Trafton (submitted) argue that theories of graph comprehension need to include an explicit spatial component in them to be comprehensive. Specifically, they argue that current models of graph comprehension account for how information is encoded only when the data are directly represented on the graph, such as a data case; they do not describe how, information is assembled when it is not directly represented in the graph, as is the case in my study when the experts apparently imagined the location of an average in a distribution that was not explicit shown. Trickett et al. videotaped several expert participants as these experts carried out
explorations of their own data, who made many reference to patterns and features not explicitly represented in the display. Trickett et al used the term “spatial transformation,” to describe the mental processes they imagined experts used to create these implicit features.

It may be that the different visual search patterns (i.e., predominantly horizontal or vertical movements) and verbal comparison methods data analyzers use may be explained by the type of spatial transformations that people want to make. Data analyzers looking to use group aggregates such as the mean to compare groups may for that reason use horizontal search patterns to locate the aggregates in the first place. Alternatively, other data analyzers who do not spatially transform the data may use slices because they only need to compare the counts of data cases and so choose to use vertical search movements. A study by Bakker (2001) suggests that children also use spatial transformations in reasoning about data. Bakker observed how 11- and 12-year old students developed their concepts of a center. In a lesson examining graphs showing the weights of students in a class of seventh-graders, the teacher described the center of the group as a “bump,” a visual characteristic of the stacked-dotplot where most of the cases were “bunched up” to form the peak or hill of the distribution. Describing the data near the mode of a distribution as a “bump” may serve to spatially transform the distribution in the manner described by Trickett et al.

Indeed, students in the Bakker (2001) study quickly adopted the construct of a “bump” and began to describe distributions in terms of their centers, where formerly they described them in terms of individual cases. They even used the idea of a “bump” to summarize other graphs, such as value bar graphs as shown in Figure 26. What is
significant here is that the students had learned to see what was a visual bump in the stacked dot plot representation as the area where the changes in bar heights flatten off in the value bar representation. This indicates that the students understood the term “bump” as a description of where the majority of the cases in a group were clustered, and not just a perceptual characteristic of a particular graph. In a subsequent lesson discussing the weights of eighth-grade versus seventh-grade students, one student said, “The [eighth-grade] bump would be more to the right” indicating that she was now viewing group differences in terms of shifts in the location of their distribution centers.

![Figure 26. An example of a value bar graph showing the weight of students. In a value bar graph where each bar represents one data case. The length of a bar indicates the numerical value, in this instance the weight of a student. Note that in this graph where the values are most dense is indicated by a “flattening” of the bars.](image)

The findings in this thesis add to the growing literature examining how novice and expert data analyzers approach data. It further documents several common, but statistically unconventional, methods that students use to compare groups, namely using local slicing and cutpoint comparison methods (cf., Gal, Rothschild, & Wagner, 1989;
Ben-Zvi & Arcavi, 2001). It also found that novices used more global comparison methods, such as global displacement (e.g., means, modal clumps) when comparing two groups with a large difference between the means.

The different ways that experts and novices visually searched graph pairs validate the notion that novices and experts think about and seek information from graphs in fundamentally different ways. While differences in how experts and novices visually approached the two-group comparison task seems related to how they verbally discussed data, this does not imply that changing the way a novice either talks about or visually inspects data will necessarily change the way he or she thinks about the data. However, these would be interesting instructional possibilities to explore.

This thesis was not designed to test methods for teaching students how to make group comparisons using aggregate methods. But at least two possible strategies are suggested by the results. The first is to introduce students initially to distributions in which there are relatively large differences between means, before moving to the more common situation where the means are somewhat closer, or nearly equal. This suggestion is based on the results showing that novices were more likely to use global comparisons methods in these instances. Perhaps after exploring and thinking about the methods they use in cases where the means are far apart, they would be more likely to use the same methods in situations where the means are closer together.

Another possible instructional intervention that could prove helpful would be to use video animation to present graph pairs to students using the sweeping motion that experts use to view each graph separately. Viewing such a simulation might help students to notice emergent properties of the distributions, such as centers, that they otherwise fail
to notice, Having noticed them, they then may be more likely to use them for purposes such as for comparing two groups. Grant and Spivey (2002) found support for such an instructional possibility when they showed that reasoning about a diagram-based problem could be improved by guiding students’ visual attention during the problem-solving task.
Finally, here are a few questions about you.

1. Age

2. Sex

3. Please circle: Are you (a) an undergraduate
   (b) a graduate student
   (c) a faculty of the psychology department
   (d) other?
   If other, please specify:

4. If still a student, please indicate:
   Major
   Year of study

5. Please tick: How much statistic experience have you had?
   - None
   - 1-2 courses
   - 3-5 courses
   - 6 or more course
   - Taught statistics for 1-5 years
   - Taught statistics for 6-10 years
   - Taught statistics for more than 11 years

6. Have you ever taken a calculus course?
   yes / no
7. Please tick: How much college mathematics experience have you had?

- None
- 1-2 courses
- 3-5 courses
- 6 or more course
- Taught mathematics

8. On a scale of 1 to 7, indicate how well you understand statistics?

(not confident at all) 1 2 3 4 5 6 7 (very confident)

9. Have you ever used a statistics computer software to analyze data?

yes / no

If yes, did you use the software to make plots such as the ones you saw today?

yes / no
APPENDIX B
RFV CODE

Below is an excerpt of the RFV code used in the experiment. The code shown is for the RFV window size, and for the stimulus called “FISH” giving the more phrased question.

// RFV code for MSC stimuli (2 October 2002)

defaults

frame_size(1024, 768)
focus_window(
  155 155
  155 155
  153 153
  150 150
)
motion_blur_speed(7000)
text_font("SansSerif", BOLD, 18)

... ...

// Context for the “FISH” graph pair. Data here are obscured unless viewed through the RFV viewing window

//FISH DATA MORE
component("N1M"

block

display

  text_line("The graph you will see shows data about the weights of fish")
  text_line("from two species, the unmod and the genmod.")
  text_line("Weight was measured in pounds.")
  text_line("Each circle in the graph represents one fish.")
  v_space(20)
  button("NEXT")

  )

block

display

  stimulus("sdp_WHITE.JPG"
           "sdp_WHITE.JPG"
           "sdp_WHITE.JPG"
           "sdp_WHITE.JPG"
           "sdp_WHITE.JPG"
           "sdp_WHITE.JPG"
          )
  v_space(20)
  text_line("Do the unmods tend to be heavier than the genmods?")

77
text_line("Please click on the NEXT button")
row(
    h_space()
    button("NEXT")
    h_space()
)
)
)

block ( 
    display( 
        stimulus( 
            "sdp_botm_GMFISH_BASE.JPG"
            "sdp_botm_GMFISH_BASE.JPG"
            "sdp_botm_GMFISH_BASE.JPG"
            "sdp_BORDER.JPG"
            "sdp_botm_GMFISH_n.JPG"
        )
    )
    v_space(20)
    text_line("Do the unmods tend to be heavier than the genmods?")
    text_line("Please click on the button to give your answer")
    row(
        h_space()
        button("YES")
        h_space()
        button("NO")
        h_space()
        button("NO, they are about the same")
        h_space()
    )
)

feedback( 
    "YES"
    time_limit(1500 strict)
    display( 
        text_line("You answered YES, they tend to be heavier")
    )
)

feedback( 
    "NO"
    time_limit(1500 strict)
    display( 
        text_line("You answered NO, they tend to be lighter")
    )
)

feedback( 
    "NO, they are about the same"
    time_limit(1500 strict)
    display( 
        text_line("You answered NO, they are about the same")
    )
)
// Instructions for the Verbal Justification trials
component("Verbal Orientation"
    block(
        display(
            text_line("Now you will go to the second part of the experiment.")
            v_space(20)
            button("NEXT")
        )
    )
    block(
        display(
            text_line("You will see the graphs you saw earlier but this time the circles
do not disappear.")
            text_line("If you think you need to change your answer from what you
gave earlier,")
            text_line("you can change your answer here.")
            text_line("Please read the questions and explain to the experimenter how
you got your answers.")
            v_space(20)
            button("NEXT")
        )
    )
)

// Presenting the FISH graph pair but without obscuring the data using the RFV

// FISH DATA MORE
component("secondN1M"
    block(
        display(
            text_line("The graph you will see shows data about the weights of fish")
            text_line("from two species, the unmod and the genmod.")
            text_line("Weight was measured in pounds.")
            text_line("Each circle in the graph represents one fish.")
            v_space(20)
            button("NEXT")
        )
    )
    block(
        display(
            stimulus("sdp_botm_GMFISH_BASE.JPG"
            "sdp_botm_GMFISH_BASE.JPG"
            "sdp_botm_GMFISH_BASE.JPG"
            "sdp BORDER.JPG"
        )
    )
)

79
Do the unmods tend to be heavier than the genmods?

Please show the experimenter how you got your answer, then click your answer.

Row:

- Button "YES"
- H-space
- Button "NO"
- H-space
- Button "NO, they are about the same"
- H-space

...
APPENDIX C
HOW TO CODE A TRANSCRIPT

Example 1:

Context provided for graph

The graph you will see shows data about the depth of top-soil specimens taken at two geological stations, the Valdai and the Rosemount.
Depth of top-soil was measured in cm.
Each circle in the graph represents the depth of one top-soil specimen.

Graph stimulus and question:

STATION

Valdai

Rosemount

DEPTH (cm)

Does the Valdai top-soil tend to be shallower than the Rosemount top-soil?
Transcript of PN024’s verbal response in the justification task

Below is the complete transcript of the justification response that this participant offered in this trial. A comparison method to code was found based on two criteria: (1) The coder was able to locate the participant’s definitive answer about whether Valdai top-soil was shallower than Rosemount top-soil. (2) The coder categorized the first comparison method used in relation to the participant’s definitive answer that fit one of the eight coding categories described in the Chapter 3.

For illustrative purposes, the comparison method and definitive answer identified in the transcript are boxed and marked with an A (for answer) and C (for comparison method) below. Notice that the comparison method identified (A1 to A5) spanned across several exchanges between the participant and experimenter but the idea of the comparison method, in this case “Local Slices” remained intact.

Note also that the comparison method was assessed to locate any reference to data in the left or right regions of the graph stimulus, R (for reference). If participants made no tail references, then no R would be located. In this case, the participant did make a tail reference and it is boxed and marked (R1 to R3). Since the participant made equal reference to the left, right, and central regions of the data, this participant was regarded as not being biased to exclusively referencing a region on the left or the right.

The number following the A, C, and R is simply a sequential marker to keep track of the answer, comparison method, and reference coded.

PN024: Um, I think [they are about the same (A)] because they are more spread out.

I: More spread out? What do you mean more spread out?

PN024: Um...

I: One of them, both of them?

PN024: They are both pretty much spread up and it looks like some of them are equal.

I: Which ones are equal?

PN024: [Like these two. (C1)]

I: The one above, between 85...?

PN024: Yeah [between 85 and 90 are equal (C2),(R1)].

I: So those two columns are equal, okay, so that is sort of helping you decide that they are about the same?

PN024: Yeah, they look like they could be equal, around, like they are just dispersed in different areas kind of.
I: So, so are you just looking at those two columns or?

PN024: Well, these two are the same (C3).

I: Which ones? Can you tell me which ones?

PN024: It is between 75 and 80 (C4), (R2).

I: Okay, the one with the four.

PN024: And these are the same between 70 and 75 (C5), (R3).

I: Okay the one with the four again and because those are the same number of dots pretty much so you are thinking they are the same?

PN024: Yeah.

I: Anything you want to add? Okay.
Example 2

Context provided for graph

The graph you will see shows data about petal lengths from two types of flowers, the Versicolor and the Verginica. Petal length was measured in mm. Each circle in the graph represents one petal.

Graph stimulus and question:

Do the Versicolors tend to be shorter than the Verginicas?
Transcript of PN029’s verbal response in the justification task

Below is complete transcript of the justification response that this participant offered in this trial. As above, the transcript is marked with boxes to indicate the participant’s definitive answer (A), comparison methods (C) and any mention of a tail reference (R).

This participant offered three comparison methods. The only one that was coded was the first comparison method offered, which was the mode. This comparison method is boxed and marked with a (C1 and C2) in the transcript below. For illustrative purposes, the second and third comparison methods are noted as (CC) and (CCC).

Note that the first comparison method was not marked with a tail reference (R) because the participant did not mention any side of the distribution in her first comparison method. This participant was coded as not mentioning any tails. Had the second (RR) or third (RRR) comparison method been coded, then this participant would have been coded assessed having made a tail reference.

PN029: Yes (A1), the Versicolors tend to be shorter (A2) than the Virginicas.

I: Because?

PN029: Because if you look at the mode for Versicolor it is at 45 (C1), and if you look at the mode for Virginica it is at 49 (C2). and again, as a general trend, Versicolors start at 30 and Virginica doesn’t even start (RR) until 42,43 (CC) and the same goes for the other end. Versicolor ends (RRR) at about 56 and Virginicas end at 66 or 67 (CCC). So the Versicolors tend to be shorter.
APPENDIX D

STACKED-DOTPILOTS OF HM DISTRIBUTIONS

Expert and novice HM in the 3 meandiff by 2 wording conditions, averaged over two trials.

Expert and novice HM averaged over twelve trials.
Expert and novice PropTimeR in the 3 mean difference by 2 wording conditions, averaged over two trials.

![Graph showing PropTimeR vs Meandiff and wording]

Expert and novice HM averaged over twelve trials.

![Graph showing PropTimeR]
REFERENCES


