[UNIVERSITY OF HUDDERSFIELD]

The interactive effect of Gestalt laws of perceptual organisation and task demands on graph comprehension

A thesis submitted to the University of Huddersfield

in partial fulfilment of the requirements for

the degree of Doctor of Philosophy

Nadia Ali

April, 2011

Copyright Statement

Acknowledgements

Firstly, I would like to express my gratitude to my supervisor, Dr David Peebles, for giving me the opportunity to do this PhD. Your enthusiasm inspired me to learn more than I thought I ever could. Your constant support, both professionally and personally will always be appreciated.

I am also indebted to Dr Dave Tyfa for his help and comments throughout which forced me to think more critically about my work. Thanks are also due to Dr Robert Ward and Dr Dave Robinson for their advice.

Finally I thank my family and friends for their love and support.

Abstract

I describe a series of seven experiments investigating how undergraduate students' comprehension of 2x2 `interaction' bar and line graphs widely used to present data from two-way factorial research designs is affected by both the graph format and the nature of the interaction with them.

The first four experiments investigate how different Gestalt principles of perceptual organization operate in the two graph formats and demonstrate the effects of these principles (both positive and negative) on graph comprehension. In particular, Gestalt principles are shown to hinder significantly students' comprehension of data presented in line graphs compared to bar graphs and that the patterns of errors displayed by students are systematic. The analysis also informs the development of two modified line graphs, one of which improves data interpretation significantly to the level of the bar graphs.

The final three experiments investigate more deeply how the processes involved in different types of interaction with graphs affect users' comprehension of the data depicted. In the first four experiments, participants attempted to understand the graphs while thinking aloud. However, a subsequent study (Experiment 5) demonstrated that writing an interpretation produced significantly higher levels of comprehension for line graphs than when thinking aloud. The final two experiments sought to identify the cause of this difference by isolating demands specific to the verbal protocol condition.

The results of this research show that (a) in certain circumstances the Gestalt principles of perceptual organization that operate in different graph formats can significantly affect the interpretation of data depicted in them but that (b) these effects can be attenuated by the nature of the interaction. The implications of this research are that identifying an appropriate method of interaction as well as ensuring appropriate display design ensures that the majority of users will be able to interpret these graphs appropriately and so recommendations can be made for graph use in educational settings.

Contents

Figures

Tables

Chapter 1. Introducing the concept of statistical literacy

Introduction

The aim of this research is to develop a theory of how people understand a form of diagrammatic representation widely used in the natural and social sciences (the 2x2 'interaction' graph) and to then apply this theory to improve people's ability to use such graphs—either through enhancing the diagram or by identifying the most appropriate form of interaction (or both).

The idea for this research arose from the experience of witnessing undergraduate psychology students struggle to interpret the results of 2x2 factorial designs in their research methods classes and to misinterpret graphical representations of the data produced by such experimental designs. Although the vast majority of undergraduate students can be expected to have worked with standard bar and line graphs representing the relationship between two or more variables, there is a limited amount of research investigating students' conceptual understanding of these graph types.

In this introductory chapter I provide a rationale for my research by outlining the importance of statistical literacy in general before considering the notion of graphical literacy, followed by a brief consideration of the unique characteristics of graphs which allow them to communicate information which can be processed rapidly and easily when the viewer is familiar with the graphical conventions present in the diagram.

Statistical literacy

In today's technologically advancing society emphasis on numerical and statistical literacy has increased to the point where it is now considered as important as the traditional notion of literacy whereby individuals are expected to be able to read and write to a conventionally acceptable standard (Gal, 2002). The ability to understand quantitative data influences important daily decisions we make, such as which school to send our children to, which university to study at, or whether to invest in a particular market (Watson and Callingham, 2003). To be considered statistically literate, individuals must be proficient in a range of numerical, quantitative and mathematical skills. A broader definition of statistical literacy was provided by Wallman (1993):

> "Statistical literacy is the ability to understand and critically evaluate statistical results that permeate our daily lives – coupled with the ability to appreciate the contributions that statistical thinking can make in public and private, professional and personal decisions" (1993. p1).

The importance of statistical literacy is evident when one considers the skills now required of people. According to the Centre for Statistical Education (CSE) and the American Statistical Association (ASA; 1994, cited by Mooney, 2002), 70% of the US workforce deals with quantitative information on a daily basis and people unable to make use of quantitative information are hindered from being productive employees, students, consumers and citizens. Research has shown that quantitative literacy is one of the major factors influencing an individual's ability to secure employment (Rivera-Batiz, 1992) and the size of the salary they are paid (Murnane, Willett, and Levy, 1995).

Mooney (2002) cites an impressive amount of research emphasizing the need for statistical literacy skills. Professional educational organizations including the National Council of Teachers of Mathematics (2000), the National Council for Social Studies (1994) and the National Council of Teachers of English in conjunction with the International Reading Association (1996) have documented the need for statistical skills within their respective disciplines. Accordingly, development of statistical skills is part of most current middle school mathematics curricula (e.g., Bolster et al., 1994; Chapin, Illingworth, Landau, Masingila, & McCracken, 1997; Charles et al., 1998).

Although statistics and data analysis are a key component of the school mathematics curriculum (National Council of Teachers of Mathematics NCTM, 2000), for many people, this represents the limit of their formal education in the topics. In addition to this, research demonstrates that students are failing to meet basic criteria set for statistical literacy skills after graduating from school. This is an especially problematic issue as the demand for statistically literate employees is increasing in the workplace and findings from research are revealing that graduates are not equipped to meet these demands (Steen, 1999).

The importance of graphical representations

Data analysis involves a heavy reliance on graphical representations (Shaughnessy, Garfield and Greer, 1996) and a core skill of statistical literacy is to be able to construct and interpret statistical information depicted in graphs (Lowe, 1993, Shaughnessy, Garfield, & Greer, 1996, Friel, Curcio and Bright, 2001). All of the widely used 'productivity' software packages (e.g., Microsoft Office, Apple's iWork, Open Office etc.) incorporate sophisticated graph production functions and constructing such diagrams in these packages is as easy as the click of a button.

In highly specialized fields, diagrams are an invaluable resource when an individual is required to interpret large amounts of complex information. For example, in the field of meteorology diagrams which forecasters employ display numerous variables which can result in a single diagram displaying tens of thousands of data points (Hoffman, 1991). Plotting this information in a single diagram or multiple diagrams condenses information and reduces cognitive load considerably, allowing experienced forecasters to be able to locate and extract information more quickly than if the information consisted of large lists of data values (see figure 1).

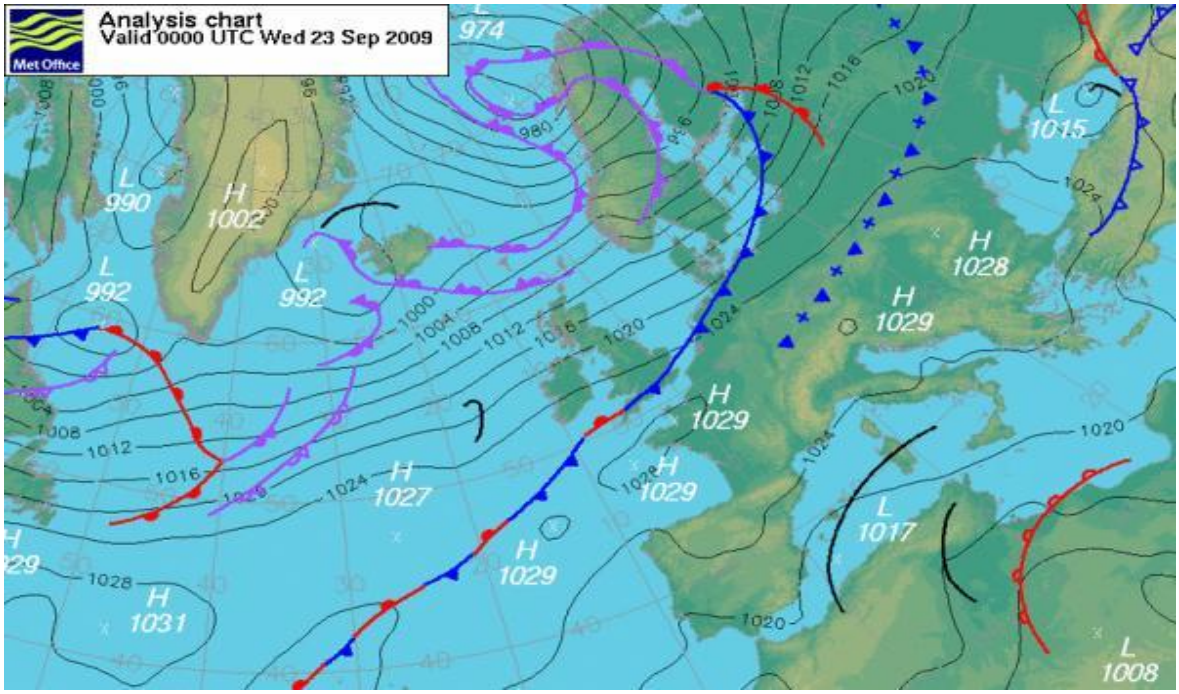


Figure 1. A surface pressure chart (Met office, 2009)

In fact, in science and social sciences, experts are often so dependent on graphs that they may be unable to do their work without them (Tabachneck-Schijf, Leonardo and Simon, 1997). In addition to this, use of graphs to depict quantitative data may be a central, possibly defining feature of science. There is a nearly perfect correlation between use of graphical displays and the "hardness" of a scientific discipline; in both the scientific research literature and textbooks (Smith, Best, Stubbs, Johnston and Archibald, 2000).

Figure 2. STOCK MARKET CHART DEPICTING BARCLAYS SHARE PRICE OVER TWO YEARS (THE Investor, 2008)

Diagrams can also be useful when imparting complex information from a particular field to a non-specialist audience. For example, the volatile nature of the stock market has received a lot of news coverage in recent years. In order to communicate the drastic change in share prices due to the economic slowdown, newspapers typically rely on graphs depicting share values over a certain period of time, so the pattern of rise and fall in share prices can be determined visually rather than the reader having to examine and compare values from a table. A glance at Figure 2 reveals that share prices for Barclays bank plummeted from February 2007 to November 2008. Therefore, use of diagrams can make complex quantitative information easy to understand.

These examples demonstrate the way in which diagrams such as graphs permeate our everyday lives and are not simply restricted to education or particular job roles. Consumers need to understand how to interpret these types of graphs because often the designer can use graphs to mislead the audience to believe a pattern exists which does not. A recent example of this which was of global interest is the BP oil spill. BP released a technical report concerning how much oil they were successfully collecting. They used a cumulative graph

rather than one based on an averages of amount of oil collected daily. Because it was a cumulative graph it appears as though the amount of oil collected is increasing and so their efforts are resulting in an improvement (see figure 3). When the graph is plotted based on amount of oil collected daily (Figure 4), a very different picture emerges concerning the successfulness of BP's efforts.

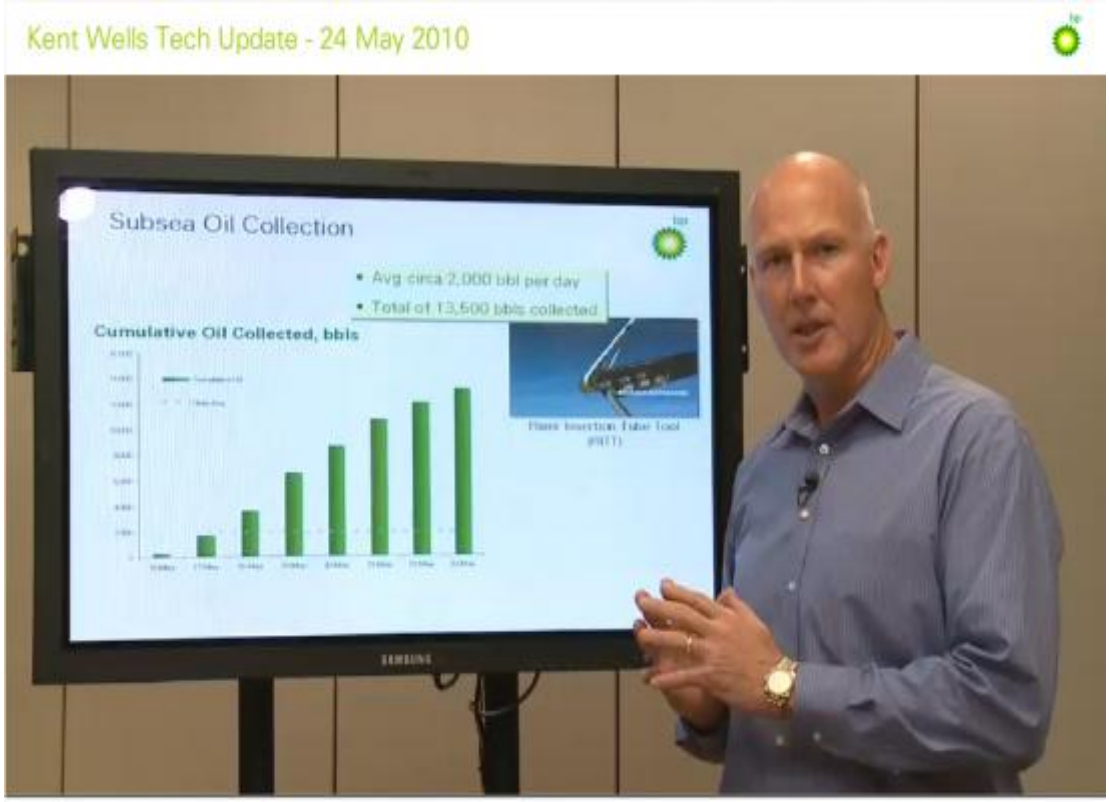


Figure 3. A CUMULATIVE BAR GRAPH PLOTTING AMOUNT OF OIL COLLECTED DAILY (BP. CONCERTS, 2010).

The line graph in Figure 4 clearly demonstrates that the amount of oil collected daily has actually dropped well below the average for amount of daily collection for this period as a whole. The BP report used a cumulative graph to place a positive slant on their collection efforts. As the presentation took the form of a technical report, graphs were used to allow consumers to more easily understand the effect BP's efforts were having on the spillage. Unfortunately, if the consumers were unaware of cumulative graphs or were not paying attention to the title they could be misled to believe that the average rate of oil being collected daily is increasing.

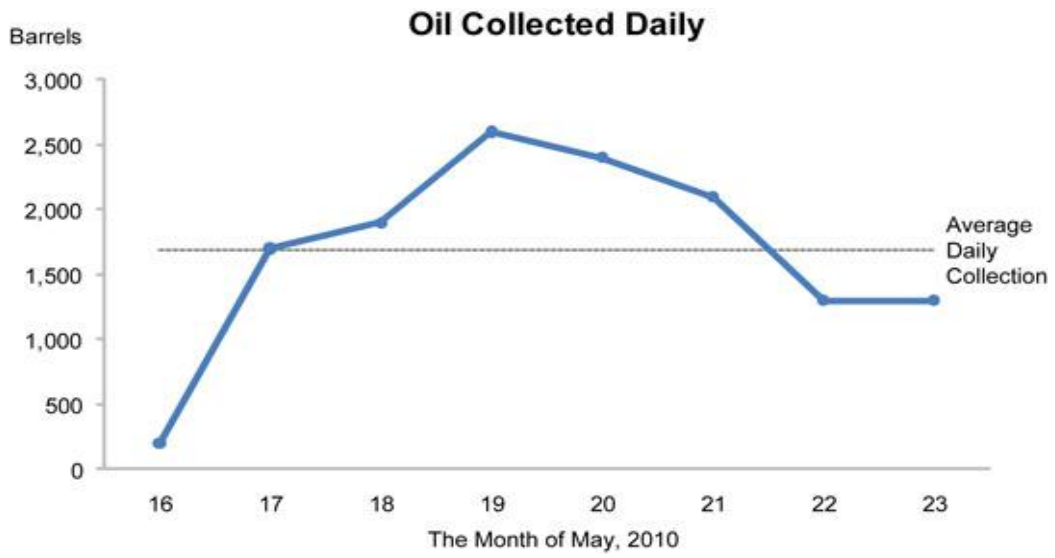


Figure 4. A line graph depicting amount of oil collected daily (flowing data, 2010).

This is a clear example of why it is important that individuals become proficient in statistics and how to read diagrams. The news of the BP oil spill was a global phenomenon which dominated the headlines for several weeks. If viewers are unable to apply their statistical knowledge to the information being communicated accurately, they are likely to be misled about what is happening in the world.

Understanding how graphs work

The increasing popularity of graphs has been attributed to the fact that they make quantitative information easy to understand (Bertin, 1983, Larkin and Simon, 1987, Pinker, 1990, Kosslyn, 1989, 2006). Although a number of graphs exist they all share a single defining property, in that they all use spatial or visual characteristics - for example, length, area, colour hue - where a change in the spatial or visual characteristic represents a change in quantity. For example, GDP across different countries can be represented by colour hue, with darker colours showing a higher GDP for wealthier countries (Bertin, 1983, Winn, 1987, Pinker, 1990).

What makes graphs unique is that the visual pattern at the centre of the display (e.g., an increasing line) can depict quantitative relationships between variables that can be identified instantly if the reader is familiar with what the pattern signifies (so in the case of an increasing line the reader would be aware that the visual pattern is depicting a relationship illustrating that as one variable increases, so does the other). Non-graphical

methods of depicting data (e.g., tables) would require cognitive effort in terms of time and mental computation to decipher the quantitative relationship if there are numerous data values to compare. This is because there is no visual pattern to indicate what the relationship is between data points so readers have to compare the data points to establish what the relationship is (Bertin, 1983, Pinker, 1990).

The power of graphical displays has long been acknowledged. Use of graphical displays dates back to ancient times – before language evolved. The earliest graphic displays to have been discovered are geographic maps etched in clay and are thought to date from the third millennium BC (Bertin, 1983). Long before communication of statistical information using graphical displays took hold innovative individuals used graphs to depict important patterns emerging from data.

A classical example is that of Florence Nightingale who changed medical practice by implementing what is considered nowadays as basic hygiene practices (e.g., changing bed sheets for each new patient admitted to hospital). The importance of the changes she made was recognized because she documented each and every change she made and what improvements resulted from those changes. These descriptive statistics were easily communicated in a graph she created, termed "polar area graph" allowing the pattern that emerged to be more easily perceived than if the data had remained in numerical form. Depicting rate of deaths from various causes using a graph allowed viewers to instantly perceive the shocking results that had emerged from data collection; soldiers were much more likely to die from preventable diseases than actual war wounds.

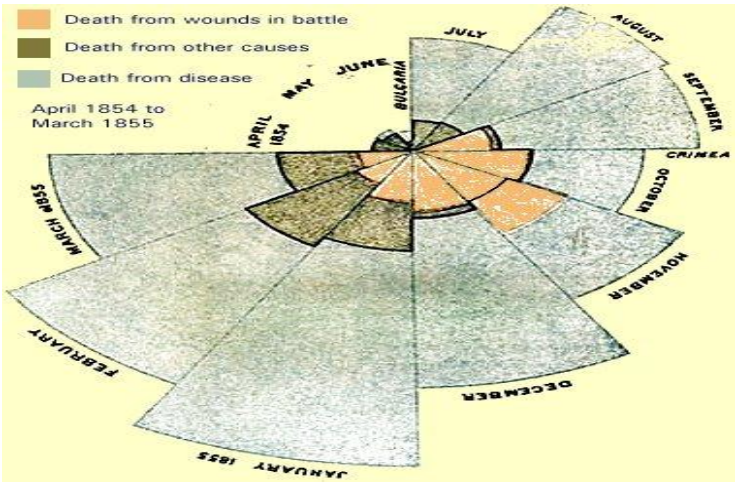


Figure 5. Causes of mortality in the army (Improving visualization, 2009).

Since then graphics have evolved and the study of the relative usefulness of graphical presentations took hold in the 19th Century. As the number and type of graphical displays have increased so have the attempts to produce a set of guidelines for which type of graphical display should be used. In 1915 the Joint Committee on Standards for Graphic Presentation emphasized the need for well thought out guidelines for consistent use of graphical displays based on the assertion that:

> "If simple and convenient standards can be found and made generally known, there will be possible a more universal use of graphic methods with a consequent gain to mankind because of the greater speed and accuracy with which complex information may be imparted and interpreted"
>
> (McCall, 1939, p. 475, cited in Friel, Curcio and Bright, 2001).

A number of graphic handbooks have been published providing recommendations on how graphs should be designed, some based on authors' intuitions, others on experimental results (Brinton, 1914; Schmid & Schmid, 1979; Tufte, 1983, Cleveland, 1985; Kosslyn, 1989, 1994, 2006). The extensive use of graphical diagrams in a wide range of contexts means that these guidelines are invaluable to anyone wanting to make a sound judgment on which graph format to use when communicating information. Some of these guidelines are considered "common sense" whereas the wisdom of others is only appreciable after the individual has had the opportunity to analyze what effect following (or not following) the guideline has on comprehension (Kosslyn and Chabris, 1993).

Although guidelines for graph construction are an invaluable tool for designing graphs that convey their message clearly, well designed graphs do not ensure that readers can comprehend the message the graphical display is supposed to be conveying. The heavy reliance on pictures to communicate information gives the impression that pictures are easier to understand than other forms of language such as verbal, written and mathematical language. However, pictures that the general public are exposed to differ greatly from diagrams used in specialist subject areas. The former relate to everyday matters which people are familiar with and are easy to understand (e.g., the icon for a fire extinguisher) (Tversky, 2001). The latter are used to convey information specific to the field they are created for. Therefore, authors have pointed out that

diagrams such as charts and graphs can be just as difficult to interpret as other methods of depicting numerical data (Vernon, 1946, Lowe, 2000).

Lowe (2000) points out those diagrams used specifically to convey quantitative and qualitative information in specialized fields are more difficult to comprehend than pictures that permeate our everyday lives. Because these diagrams are unique to specific fields, individuals will not learn to interpret them due to exposure in their everyday environment the way they learn to interpret visualizations that are not specific to a specialist field. Instead, students and employees need to be explicitly taught how to interpret these types of diagrams so they can successfully reason with and translate information depicted in them.

Lowe (2000) suggests that part of the reason these diagrams are more complex than diagrams used in everyday contexts is because they are unfamiliar to everyone except specialists in the field. Furthermore, unlike pictures that permeate our everyday lives visualizations in specialist fields are not meant to be taken literally. Rather, these types of diagrams use a number of graphic conventions to depict relationships between variables which readers need to be familiar with in order to extract the information required from the diagram.

In relation to the diagrams used in the experiments for this study, all of the above points are relevant. Graphs used to present the results of two way factorial experiments – often referred to as *ANOVA* or interaction graphs - are commonly encountered in the natural and social sciences but are rarely encountered outside these fields and so are only familiar to specialists in the disciplines. Readers must be familiar with the graphic conventions employed in these graphs in order to correctly interpret them.

The notion that diagrams such as graphs used in specific fields can be difficult to interpret has been demonstrated in the graph comprehension literature. Vernon (1946) found that when non-educated participants were asked to interpret graphs they struggled and often described the appearance of the graph rather than extracting the quantitative relationships the graphs were depicting. Furthermore, with certain graph types some participants were unable to extract the general and most obvious trend the graph was showing. Vernon (1946) also found that some responses were based on personal beliefs rather than information depicted in the diagrams presented to participants. She concluded that if individuals do not

possess the necessary skills required to interpret visualizations then displaying information in charts, graphs and pictorial displays will not be of any use and can be misleading.

It is not only uneducated participants who struggle to read or construct graphs correctly. Research in the domain of mathematics has revealed that students struggle to understand mathematical relationships expressed in graphs or construct appropriate graphs (Paulos, 1988, Carlson, Jacobs, Coe, Larsen, & Hsu, 2002). This is even the case if students are high performing calculus students. Carlson et al (2002) investigated students ability to accurately construct and interpret graphs demonstrating rate of change between two variables (e.g., as more water is added height of line increases). He found that when A grade calculus students were instructed to construct a graph of a particular problem only 25% of students constructed one that appropriately represented the correct solution. In addition to this, they found that students made simple errors such as assuming the Y variable was the independent variable and the X the dependent variable.

Based on these previous studies a reasonable conclusion would be that certain types of visualizations such as graphs are not easy to comprehend unless the intended audience is trained in the skills required to be able to meaningfully interpret the information depicted in them. This assumption was further reinforced by the results of an initial pilot study conducted by Peebles and Ali (2009). They found that when asked to interpret three-variable interaction graphs a large proportion of participants were unable to extract basic level information from the graph when the data were plotted in a line graph format. Similar to Vernon (1946) they found that participants tended to focus on the appearance of the graph rather than extracting the relationship the graph was showing. When instructed to try and interpret the relationship the graph was showing many participants struggled and often missed crucial information (e.g., one of the variables) or misinterpreted information depicted in the display.

To conclude, visual imagery such as diagrams, graphs and maps permeate our everyday lives. The reliance on visualizations to communicate information is based on a dearth of research demonstrating that pictures have a number of advantages over words when communicating information (Tversky, 2001). However, as Lowe (2000) points out, although diagrams used in specific fields to communicate quantitative information

have a number of advantages over other forms of presenting data this does not guarantee students will be able to easily understand the patterns found in the data. Therefore, there is a need for research investigating how students understand graphs within their respective field.

Overview of the thesis

Three sets of studies examine (1) students' conceptual understanding of two informationally equivalent graphs as a function of display design (2) how the nature of the interaction affects users' interpretations and (3) the appropriateness of the verbal protocol method as a research methodology. In these studies participants were asked to interpret three-variable bar and line graphs used to display the results of 2x2 factorial designs in the social sciences.

The first set of studies examines the interpretation provided by graph users when viewing bar or line graph displays. An analysis of the pattern of errors reveals these two graph formats, although informationally equivalent, do not result in the same patterns of interpretation. Specifically, the limitations that influence a viewer's comprehension of line graphs are not mirrored in bar graph displays. Overall, the results suggest that comprehension of graphs is dependent on the number of Gestalt principles allowing for successful association of pattern to referents in the display.

An error analysis framework developed from the results of the experiments indicates the Gestalt principle of similarity allows for the successful association of lines at the centre of the display to variables plotted in the legend. However, due to no associative perceptual feature allowing for the successful association of lines to the variables plotted on the x axis, the majority of participants in this study were unable to interpret relationships depicted in line graphs at an elementary level. This imbalance in perceptual features allowing viewers to successfully associate pattern to variables is not present in the bar graph display. The Gestalt principle of similarity is present allowing viewers to associate bars to legend values and the bars are rooted to the values of the x axis, thus creating a balanced representation allowing for the successful association of pattern to all the variables.

This analysis led to the development of two modified line graph displays which were designed to overcome the limitations that constrain novices' interpretations of the standard line graphs prevalent in the literature.

Results revealed incorporating the same Gestalt principle allowing successful association of lines to legend values to allow association to x axis values redressed the imbalance found in standard line graphs. These findings provide strong evidence that Gestalt principles can be employed to improve base level comprehension of relationships depicted in statistical graphs.

Experiment 5 investigated whether altering the nature of the interaction with the graphs would affect performance. Performance was compared in two conditions, one where participants were required to think aloud and another where they provided written responses. Results revealed performance was superior in the written condition and the bar-line difference found in previous experiments where the think aloud method was employed was not present in the written condition. In this condition participants provided similar interpretations for both bar and line graphs and were significantly less likely to provide an erroneous interpretation in the line graph condition.

These findings led to another research question investigating which process in the think aloud condition results in the poor performance found in earlier experiments. Three ways in which this method differs to others were identified and investigated to determine whether it was some feature of the verbal protocol method which resulted in the poor performance observed

Firstly, in this method the experimenter is present and participants are required to think aloud in their presence whereas with other methods this is not the case. To determine whether experimenter presence were affecting participants' performance Experiment 6 involved participants thinking aloud without the experimenter present. Results revealed no improvement in performance when compared to conditions where the experimenter was present. Therefore, this explanation cannot account for the performance differences observed when employing the two different methods.

Secondly, this method requires participants to think aloud throughout the task whereas with other methods this additional requirement is not present. To determine whether demands of verbalization was affecting performance Experiment 7a involved participants reading the graphs silently then once they felt they had understood it stating their interpretation to the experimenter. Results were consistent with the written condition, performance in this condition was superior to that of the think aloud condition. These findings

provided strong evidence that the demands of verbalization interfere with processes involved in comprehension and result in the pattern of performance observed in conditions where participants were required to think aloud.

The final experiment acted as a control condition to Experiment 7a. To determine whether it was the act of communicating understanding to the experimenter which resulted in the observed performance improvements – rather than the explanation proposed which is that the requirement to verbalize interfered with processes involved in comprehension – this experiment required participants to think aloud throughout the task and then also provide the experimenter with their interpretation. Results revealed no improvement in performance when compared to the standard think aloud conditions. These experiments provide evidence for the conclusion that verbalizing thought processes interfere with cognitive processes involved in comprehension of material.

Based on the results of these experiments I conclude by making recommendations for graph design, the nature of interaction users would benefit from and considering whether the verbal protocol method is appropriate to employ for certain types of research questions. The next section reviews relevant literature in the area of graph comprehension which has contributed to our understanding of the factors which interact to influence graph comprehension.

Chapter 2. Research on graph comprehension

Introduction

This chapter will review research in the area of graph comprehension that has led to our understanding of the numerous factors that interact to determine how well a graph will be understood by the reader. The review will consider major contributions to the area of graph comprehension which have advanced our understanding of how graphs are processed by a reader and the limited scope of empirical research when the question of interest is comprehension of statistical graphs.

Firstly, the processes involved in graph comprehension will be considered and research relating to these processes will be briefly outlined. Secondly, the three major factors identified in the literature as influencing graph comprehension – graph format, task requirements and reader characteristics - are discussed as comprehension cannot be understood without considering how these three factors interact to shape the type and quality of information that is extracted from a graph. These three intertwining factors will then be considered in the predictions Pinker (1990) makes in his model of graph comprehension which is accepted as providing a sound explanation of the cognitive processes involved in graph comprehension (Lewandowsky and Behrens, 1999).

Carpenter and Shah's (1998) model will also be discussed because the emphasis is on within-context graphs – where the axes of graphs are labelled with meaningful variables and a title is included providing a context for the relationship depicted. Furthermore, their research focused on three-variable interaction graphs, similar to the stimuli used for the experiments reported here. Although there are numerous models of graph comprehension in the literature these two are the most relevant to the study of the types of graphs used as part of this research project to investigate graph comprehension. Finally, the research conducted for this project will be outlined and the rationale for the research project will be introduced.

The three main processes involved in graph comprehension

Theories of graph comprehension have identified three major processes involved. The first key process involves readers identifying the major visual patterns. In the case of line graphs the reader may encode the number of lines, whether they are increasing, decreasing, constant and so on. In the case of bar graphs features that may be encoded are number of bars, height of bars in relation to each other and whether they are tall or short, etc. (Bertin, 1983, Pinker, 1990).

The second process requires readers to relate the visual pattern to known quantitative trends stored in long term memory (Pinker, 1991). When the visual pattern a graph is depicting is associated with known quantitative trends (e.g., a horizontal line means that the independent variable is having no effect on the dependent variable) comprehending the trends a graph is depicting is relatively automatic and effortless. However, if the visual pattern is not associated with known trends then comprehending the trends a graph is depicting becomes effortful and usually error prone (Pinker, 1991, Shah and Carpenter, 1995).

Line graphs are considered superior to bar graphs when depicting interaction data sets because expert readers can retrieve quantitative relations from long term memory from the pattern the graph is depicting, whereas bars do not form patterns and so more cognitive effort is required to determine the trend the graph is depicting (Pinker, 1990, Kosslyn, 2006). However, novices are unlikely to have learnt these patterns and so this advantage will be of little use to them (Peebles and Ali, 2009). Therefore, bar and line graphs can be considered to be relatively equal in level of difficulty for this process when novices are required to interpret the display.

The final process is relating the visual pattern at the centre of the display to variables plotted on the axes and in the legend. Although this process initially received very little attention, Carpenter and Shah's (1998) research demonstrated that the majority of readers' time is spent reading and re-reading variable names, labels and scales on axes when interpreting three-variable graphs, thus demonstrating that this process plays a key role in graph comprehension, contrary to the assumptions of previous models (Pinker, 1990).

One of the assumptions of this research is that this final process is the most difficult for novices to perform correctly, more so in the line graph condition than the bar graph condition. The differences in design features

between graph formats means that it is much easier for novice readers to relate the pattern to variables in the bar graph condition than the line graph condition. This is because there is a salient perceptual feature depicting each level of each independent variable in the bar graph condition, a bar rooted to the x axis to allow association of pattern to x labels and colour to allow association to z labels. In the line graph condition there is a salient perceptual feature depicting the variables in the legend – lines are coloured so the pattern can be matched to z labels via a simple process of colour matching - but not the variables plotted on the x axis (Peebles and Ali, 2009).

Differences in display design have already been shown to have a striking effect on whether participants were able to comprehend interaction graphs at an elementary level (Peebles and Ali, 2009). An initial experiment investigating participants' level of comprehension when asked to spontaneously interpret three-variable interaction graphs presented in either a line or a bar graph format found that 39% of participants in the line graph condition could not be categorized as possessing elementary level graphical skills due to the mistakes they made when attempting to interpret the graphs. Conversely, no participants were classified as being pre-elementary in level of comprehension in the bar graph condition. As the graphs were informationally equivalent, differences in comprehension can be attributed to the differences in graphical pattern at the centre of the display.

The three main factors influencing graph comprehension

Graph comprehension is dependent on three intertwining factors that influence readers' ability to execute the processes involved in graph comprehension: graph format, task requirements and reader characteristics. When assessing comprehension of different diagrams all three factors need to be taken into consideration as they will interact to influence the outcome (Gray & Altmann, 2001). The literature that has contributed to our understanding of how these factors influence graph comprehension is reviewed below.

Graph format

One line of research into graph comprehension is concerned with how accurate viewers are at judging quantities from different graphical displays (typically simple, unlabelled graphs to determine how the visual pattern affects speed and accuracy of judgments). The aim of such research was to establish guidelines suggesting which graph format(s) should be employed to portray statistical information. For example,

initially, early research compared bar and pie charts to determine which were superior for accuracy of judgments. Conflicting results emerged from these experiments with some experiments revealing bar charts were superior in producing accurate judgments whereas other experimental findings demonstrated pie charts were superior to bar charts (Brinton, 1914; Ells, 1926; Croxton, 1927; Croxton & Stryker, 1927; Von Huhn, 1927; Croxton & Stein, 1932). These conflicting results revealed a simple comparison of different graph types was not an exhaustive analysis of how accuracy of judgments is affected by graph format.

This line of research was considered too simplistic and research comparing various graph formats was developed into a theory of graphical perception (Cleveland and McGill, 1984). This approach identified graphical perception tasks a person engages in when extracting quantitative information from common graphs and then ordered the tasks according to how difficult they are to perform (illustrated in Table 1). Both experimental evidence from research findings and theory from visual perception research informed how accurate different perceptual judgments would be. Findings indicated some perceptual judgments are more difficult to perform than others and so ease of information extraction from a graph depended on which perceptual judgment a graph required. For example, position along a common scale was ranked as most accurate whereas making angle judgments was ranked as the third most accurate. Based on this ranking bar charts would be superior to pie graphs for extracting quantitative information because bar charts require readers to judge quantities from a common aligned scale whereas pie graphs require angle judgments.

The results of the experiments conducted in this area supported ordering of perceptual judgments. Consistent with predictions, judging quantities from a common aligned scale resulted in more accurate judgments than judging from a non-aligned scale or from angle but all were superior to judgments based on volume or area. Based on their findings and findings from previous research Cleveland and McGill made recommendations for graph use:

> "The theory provides a guideline for effective graph construction: Graphs should employ elementary tasks as high in the ordering as possible." Cleveland and McGill (1984, p. 531)

Therefore, graphs which require readers to make perceptual judgments from position along a common scale (e.g., bar charts, line graphs) are preferable to graphs that require perceptual judgments based on angle (e.g.,

pie graphs). Since data plotted in a pie graph can also be plotted in a bar chart the latter format should be used to illustrate quantitative relationships.

Table 1. Ordering of Perceptual tasks (from most to least accurate).

Rank	Elementary perceptual task
1	Position along a common scale
2	Position along non-aligned scales
3	Length
4	Angle, Slope
5	Area
6	Volume, Density, Colour Saturation
7	Colour hue

This theory of graphical perception was an important contribution to understanding how perceptual elements affect early stage processing of graphs, but such classifications by themselves are not an adequate account of how accurate individuals are at extracting quantitative information from graphs. Later research demonstrated that differences in accuracy between some of the judgment tasks were not as distinct as proposed by theories of graphical perception. Carswell (1992) found there was not much difference in accuracy between the position, length and angle perceptual judgments but that area and volume were less accurate than other judgments.

More importantly, research in this area focussed exclusively on perceptual processing and recommendations for graph design were based solely on results of psychophysical experiments. Numerous other factors interact with early stage perceptual processing to determine whether a particular graph format will be superior to another for extraction of quantitative information and these factors need to be taken into consideration to inform graph use.

Task requirements

The above line of research was criticized for focusing solely on how perceptual elements affect graph comprehension. Further research revealed task requirements interact with perceptual processing to determine accuracy of judgments made by participants when extracting quantitative information from graphs (Simkin and Hastie, 1987; Carswell and Wickens, 1987; Wickens and Carswell, 1995). One set of studies demonstrated this by asking participants to spontaneously interpret bar charts and pie graphs. Findings revealed when interpreting bar charts participants predominantly made comparison judgments (i.e. comparing the values of the bars) whereas if they were given pie charts to interpret they predominantly made proportion judgments (i.e. comparing individual slices with the whole).

Simkin and Hastie (1987) criticized research focusing on comparison of graph formats (e.g., Cleveland and McGill, 1984), pointing out results were confounded in these experiments as the task given to participants consistently and more importantly *only* required them to make comparison judgments for different graph formats (e.g., bar charts and pie graphs). Simkin and Hastie (1987) did the same and found that the results corroborated previous findings, both simple bar charts (position along a common scale) and divided bar charts (position along a non-aligned scale) were superior to pie graphs (angle).

However, they then asked participants to make proportion judgments for the same three graph types and found that the results contradicted those of earlier research; pie graphs were superior to divided bar charts and equal to simple bar charts. The authors concluded that display design interacts with task requirements to influence accuracy of judgments. Graphs employing position along a common aligned scale are superior for comparison judgments whereas other judgments such as proportion judgments do not necessarily benefit from a common aligned scale.

The notion that graph format interacts with task requirements to affect judgment was further strengthened with the proposal of the *proximity compatibility principle* which focuses on the relationship between processing proximity and display proximity. Carswell and Wickens (1987) conducted a meta-analysis of studies comparing a variety of graph formats that ranged from being integrative (e.g., line graphs) to separable (e.g., bar graphs). The premise of the proximity compatibility principle is that the decision of which graph format to use to present data should depend on task requirements. If the task requires

information to be integrated together processing proximity is high, in which case display proximity should also be high. This is achieved by various means, for example, data needing to be integrated could be closer together in space or share the same colour (Carswell and Wickens, 1987, Wickens and Carswell, 1995).

Perceptual proximity should be high if data integration is required because people are better at integrating information from multiple sources if the visual display is designed to encourage parallel processing rather than serial processing of information. Object displays such as line graphs encourage parallel processing as a single feature (i.e. the line) can display multiple data points so different sets of data are processed together rather than separately. This reduces cognitive demand on working memory and so a more integrative interpretation of data is facilitated. The reverse is thought to be true for bar graphs as each data point is represented by separate bars. As data points are not grouped together into a single feature people have to look at each set of data points separately. Therefore, this graph format is appropriate for tasks requiring low processing proximity, such as identification of specific data points.

These findings have led to the conclusion that configural displays are a superior form of presenting information when information from multiple sources needs to be incorporated into the task a person is trying to complete, as these types of diagrams facilitate such a process. The proximity compatibility principle proposes guidelines for graph construction: display proximity is increased, or its cost decreased, as the task integration requirements are increased (Carswell and Wickens, 1987, Wickens and Carswell, 1995).

As empirical work has consistently shown that configural displays facilitate information integration and separable displays facilitate focussed attention the proximity compatibility principle has been adopted in the construction of graphically presented data. The work done by Carswell and Wickens (1987) and Wickens and Carswell (1995) further demonstrated that task requirements interact with graph format for a range of different graphs thus extending and integrating earlier research findings concerning the interaction of graph format and task requirements. This line of research also moved the focus beyond measures such as accuracy and speed of judgments (Cleveland and McGill, 1984, Simkin and Hastie, 1987) to consider how task requirements interact with graph format for tasks graphs are more likely to be used for such as identifying trends, locating specific information and identifying extreme cases.

Reader characteristics

Another factor which interacts with graph format and task requirements is reader characteristics. Various researchers have discussed what factors determine whether a person will be good or poor at reading graphs (Pinker, 1990, Maichle, 1994, Berg and Phillips, 1994). A core component of Pinker's (1990) theory is that individuals possess general graph schemas, and when presented with a particular graph, they will create a specific schema for that graph type from their existing knowledge of what graphs are for.

Pinker distinguishes between automatic, relatively effortless "look up" processes and time consuming, inferential top-down processes. The more an individual has to use top down processes, the more difficult it will be for them to interpret graphs. If an individual's schema contains message flags allowing them to translate the visual description into quantitative information required for the task (e.g., knowing that a series of bars where the next one along is taller than the previous one visually depicts that as one variable increases (x) so does the other (y)), this information can effortlessly be extracted when the message flag is activated. However, if an individual's schema lacks necessary message flags allowing them to effortlessly translate the visual description then they will have to use inferential processes to identify the trend by looking at each bar in turn and noting differences in quantitative values.

Although not extensively investigated, some researchers have empirically investigated the relationship between reader characteristics and graph competence. Maichle (1994) found that high school students who obtained a high score on graph competence tests were more likely to extract more complex information than those who obtained a low score, in less time and with less mental effort. Furthermore, these individuals' interpretations and responses to questions were more strategic than those who were classified as poor graph readers.

Some researchers have examined the relationship between cognitive development and graph competence. Berg and Phillips (1994) investigated the relationship between 7th, 9th, and 11th graders' logical thinking and graph competence. They found that students who scored higher on logical thinking measures demonstrated a higher level of graph competence compared to students who scored lower on such measures. The conclusion from this type of research is that graphical literacy is dependent on cognitive ability.

However, this view has been challenged by Roth and McGinn (1997) who argue graphical literacy should be explained in terms of opportunity to practice the skill rather than lack of cognitive or mathematical skills. Their research employed university science and mathematics graduates and results demonstrated how expertise in diagrammatic reasoning was limited to graphs individuals frequently used and this expertise did not generalize to other types of diagrams, thus demonstrating that practice with specific diagrams improves graph comprehension. Based on the evidence it would be reasonable to assume that graphical literacy is dependent on an interaction between cognitive ability (Pinker, 1990, Maichle, 1994, Shah and Carpenter, 1995) and experience of interpreting and constructing graphs (Larkin, 1989, Anzai, 1991, Roth and McGinn, 1997).

Within-graph differences

Comprehension of the quantitative relationships a graph is depicting not only depends on an interaction between task requirements, graph format and reader characteristics but also depends on the way a graph is constructed. For example, a line graph depicting the relationship between three variables can only depict one set of trends optimally, namely the x-y trend is retrieved automatically if an individual has learnt the associations between certain patterns and the quantitative relationships they depict (e.g., a flat line is showing that the independent variable is having no effect on the dependent variable). Conversely, the relationship between the legend (sometimes referred to as z) and y values has to be inferred by comparing one level of the x value against each z value to determine the effect the third variable is having on the dependent variable (Pinker, 1990). Therefore, retrieval of x-y trends requires relatively effortless "look- up" processes, whereas retrieval of z-y trends requires inferential top down processes.

This assumption was tested by Shah and Carpenter (1995) who designed a set of graphs depicting the same data set from alternative perspectives. They found that university students gave differing interpretations to the two sets of graphs depending on the perspective they saw. Those who saw 6a typically described the effect of age on vocabulary score with minimal information concerning the effect of hours of TV watched. Conversely, those who saw 6b typically described the effect of number of hours of TV watched on vocabulary scores, with minimal information concerning the effect of age, demonstrating that a readers

understanding of x-y relations is more comprehensive than their understanding of the z-y relations depicted in line graphs

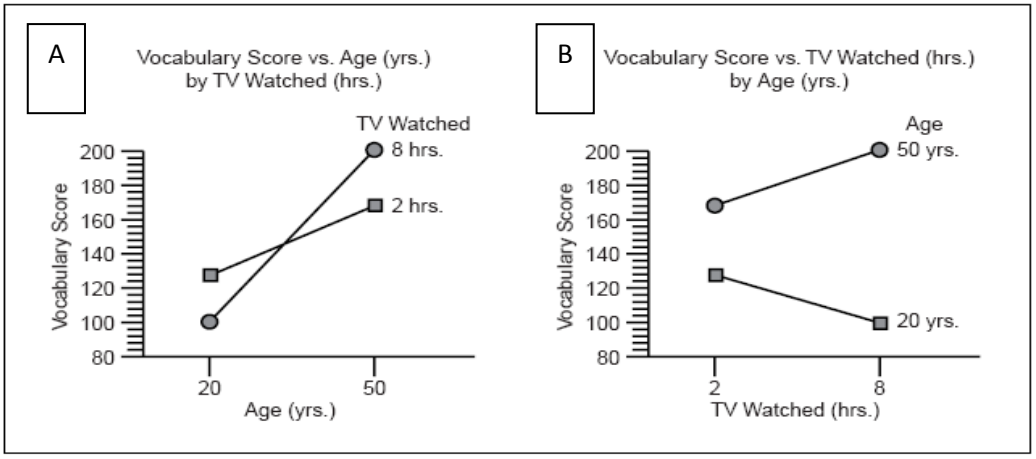


Figure 6. Differing perspectives of the same data set (Shah and Carpenter, 1995).

When individuals were asked to judge whether two line graphs depicted the same or different results participants erroneously judged the data sets to be different on 50% of the trials. Based on the results of their research Shah and Carpenter (1995) recommended that the relationship the designer wishes to communicate should be plotted on the x axis, so the quantitative relationship can be retrieved automatically. Therefore, although line graphs are superior to bar charts for depicting the interactive effect of two variables on a third, only the x-y relationship will automatically be retrieved whereas the z-y relationship will have to be inferred (Pinker, 1990). This will result in a limited understanding of the relationship depicted if novices are interpreting the graphs (Shah and Carpenter, 1995).

Shah and Carpenter (1995) investigated conceptual understanding of line graphs with more than one experimental variable. This initial research revealed limitations in comprehension of graphs depicting three variable relationships and provided valuable insight concerning how line graphs depicting statistical information should be plotted. However, their research focussed very narrowly on line graphs leaving open the question whether other graph formats plotting three-variable relationships would share the same limitations as the line graph display. In an attempt to address this question Peebles and Ali (2009) investigated comprehension of both bar and line graphs depicting a relationship between three variables. The results of their experiment demonstrated that the effect Shah and Carpenter (1995) found was reversed when

data sets are plotted in bar graphs. Using a modified coding scheme designed by Shah and Carpenter (1995) verbal statements were classified according to whether quantitative information about the x-y or z-y relationship was extracted.

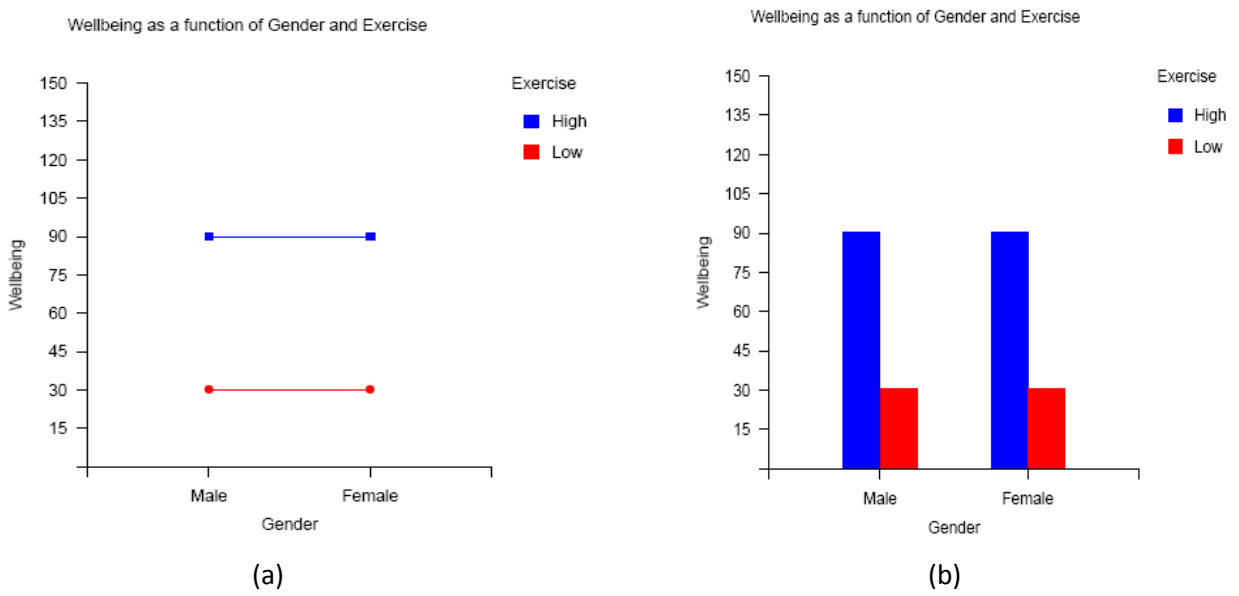


Figure 7. Graph stimuli from the experiment by Peebles and Ali (2009).

Consistent with Shah and Carpenter's (1995) results, in the line graph condition participants' understanding of x-y relations was more comprehensive than their understanding of the z-y relationship. However, this effect reversed in the bar graph condition and results showed that participants' understanding of the z-y relationship was more comprehensive than the x-y relationship. Based on the results of their research Peebles and Ali (2009) recommended that, in the case of bar graphs, the relationship the designer wishes to communicate should be plotted on the z axis so the quantitative relationship can be retrieved automatically.

This reversal effect can be explained by Gestalt principles of perceptual organization. In the case of line graphs, data points are connected by the x-y lines and so the Gestalt principle of connectedness (Palmer and Rock, 1994) means that each line will form a visual chunk. This will lead participants to use the values in the legend as a label and describe the relationship between the variables on the x-y axis. However, in the case of bar graphs the variables in the legend are grouped together by bars on the x axis and so the Gestalt principle of proximity (Wertheimer, 1938) means the cluster of bars form a visual chunk. This will lead

participants to use the variable on the x axis as a label and describe the relationship between the variables on the z-y axis.

To illustrate how viewers' understanding of data depicted in these two graph formats may differ, a typical interpretation of each graph is provided. When asked to interpret Figure 7a participants typically provide quantitative information concerning the variables plotted on the x axis using legend values as labels:

> "Wellbeing is higher when men and women do a high amount of exercise. Well being is lower when men and women do a low amount of exercise. Wellbeing is identical for males and females".

Conversely, when providing a description of the relationship depicted in Figure 7b participants typically provide quantitative information concerning the variables plotted in the legend using x values as labels:

> "Males who do more exercise have better well being than those who do low exercise. Females who do more exercise have better well being than those who do low exercise. Wellbeing is identical for high and low exercise."

Therefore, the two graphs that can depict three-variable interaction data sets both share limitations in that one trend is easily retrievable, whereas another is not and will require inferential processes to work out the effect the independent variable is having on the dependent variable. Since the most common way of presenting three-variable data sets is to plot them in a bar or line graph where the third variable is placed in a legend or to label bars and lines, the only way to circumvent the poor comprehension of the secondary trend the graph is depicting is to teach students how to interpret both sets of relationships. With sufficient training the limitations that constrain novices' comprehension of multidimensional data (Pinker, 1990, Shah and Carpenter, 1995) can be overcome.

Models of graph comprehension

Pinker's model

In a broader analysis of graph comprehension than previous approaches, Pinker (1990) proposed a model of graph comprehension that systematically explains the processes involved in graph comprehension and the knowledge available to an individual when interpreting a graph. The model is based on the assumption that graph comprehension, unlike language production, is not reliant on special purpose mental faculties. Pinker (1990) points out that graphs are a relatively recent creation and if they are particularly efficient at communicating quantitative information then it is because they make good use of humans' cognitive and perceptual mechanisms. Pinker proposed a construct called a graph schema; a knowledge structure in long-term memory consisting of knowledge of graphs with "slots" or parameters for unknown information. He argued that individuals possess general graph schemas and they create a specific schema when they encounter a new graph type from the knowledge available to them about what graphs are for and how they communicate information. This schema allows them to identify a graph as a certain type (e.g., bar graphs, line graphs, pie graphs) and then directs the search for desired information.

The model distinguishes between relatively automatic, effortless "look up" processes and more difficult, slower top-down processes. The extent to which a reader will have to use look up or top down processes depends on reader characteristics or the interaction between task requirements and graph format. If the graph format is suited to the requirements of the task then an experienced graph reader can easily extract the necessary information from the diagram. However, if the graph format is not suited to the requirements of the task then extracting necessary information requires more resource consuming processes as the diagram does not facilitate extraction of information relevant to the question.

This assertion is easily verified when attempting to determine the interactive effect of two variables on a third with reference to the graphs below. The line graph allows you to visually determine whether an interaction exists (because of non-parallel lines) whereas this would be a lot more difficult to ascertain from a bar graph. This is because a bar graph requires the reader to mentally connect the tops of the bars to determine an interaction; a difficult task because it requires them to keep the heights of all the bars in working memory. If a poor graph reader is presented with the display then graph format will not make a

difference because the message flags in their schema will not contain an entry specifying that the non-parallel lines indicate that V1 and V3 interact to influence V2.

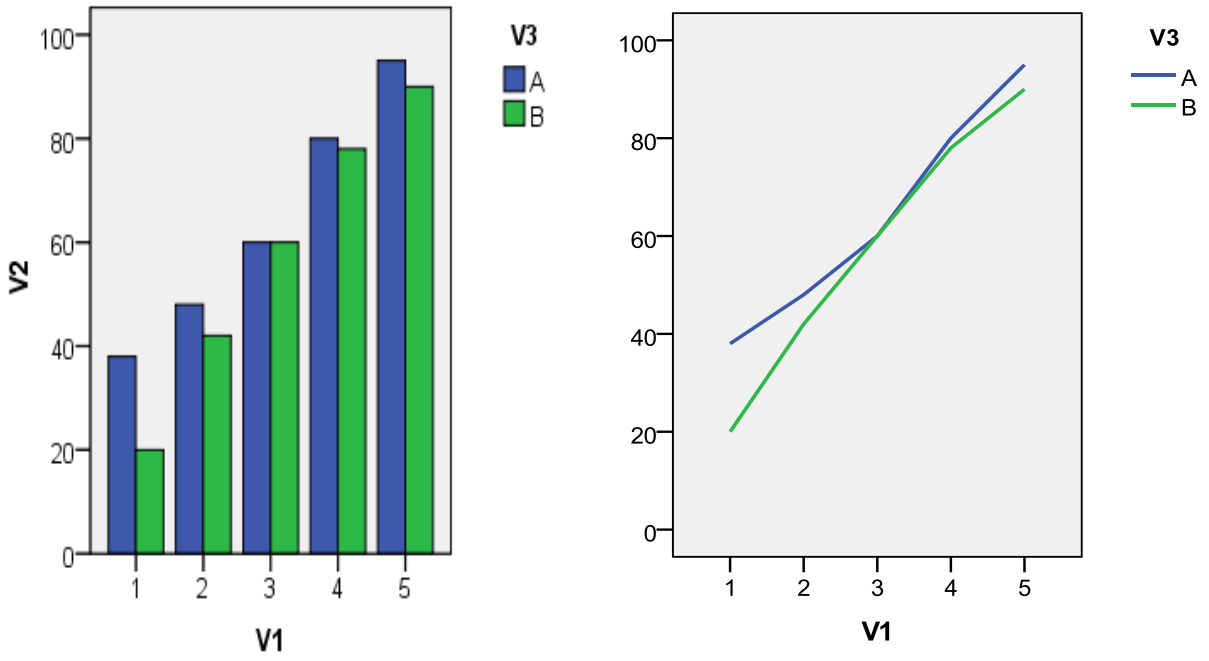


Figure 8. A bar and line graph illustrating the interaction between three variables (adapted from Pinker, 1990).

A major strength of Pinker's theory is that it considers the interaction between the three interacting elements that influence graph comprehension. As well as considering individual differences in graph reading ability his model accounts for the interaction between graph format and task requirements, thus accounting for the three main factors which interact to influence graph comprehension. The model also makes testable predictions concerning how easy or difficult it will be to interpret graphs based on an interaction of those three factors and distinguishes between automatic, effortless look up processes and inferential, effortful top down processes. Graph comprehension will be facilitated when the number of top down processes is minimized and the number of look up processes increased. This assumption has been supported in the literature (Parkin, 1983, cited by Pinker, 1990, Kosslyn, 1989, 1994, Shah and Carpenter, 1995, Shah Meyer and Hegarty, 1999).

Furthermore, the model accounts for bottom up processes by incorporating the notion of a graph schema which can explain how graphs are recognized and how attention is directed to relevant information if an individual has a strong graph schema. Although previous research had stated that reader characteristics can influence graph comprehension Pinker (1990) outlined what a graph schema may consist of and attempted to explain systematically how it can influence graph comprehension by applying top down knowledge.

Although the notion of a graph schema is a theoretical construct, some empirical studies have been conducted investigating how graph schemas can affect graph comprehension and what features of the graphical display result in a class of graphs sharing the same schema, thus providing empirical support for the existence of such a construct. Maichle (1994) investigated students' interpretations of line graphs and hypothesized that students with high scores on graph comprehension tests would possess a specific line graph schema whereas those with a low score would most likely possess a general graph schema. Students who possess a specific line graph schema should be capable of extracting more complex information and should be able to complete their interpretations and answer questions in less time and with less mental effort than students who possess a general schema. Experimental results supported these predictions.

More importantly Maichle (1994) predicted that individuals with a specific graph schema would find it more difficult to extract specific information (point reading) than trend information because trend information should be activated if individuals possess a specific schema for line graphs, thus making it difficult to isolate particular data values (Pinker, 1990). This hypothesis was supported; good graph readers verbalized three times as much (the measure used to assess mental effort) when required to verify statements involving comparison of individual point values than when comparing qualitative trends than the poor graph readers. Maichle (1994) concluded that identifying trends is an automatic process for good graph readers, thus supporting Pinker's assumption that good graph readers have the knowledge to retrieve quantitative relations stored in long term memory from the visual pattern the graph is depicting.

Miachle's (1994) study also supported the notion that graph schemas direct the search for necessary information. Analysis of participants' verbal protocols revealed that good graph readers and poor graph readers differed in the way they extracted information. Good graph readers started with an orientation phase where they identified the various variables, units and range of values and some even extracted trend

information before proceeding to read the questions and answer them. Although poor graph readers also started with an orientation phase, they spent less time in this phase and identified fewer graphical referents than good graph readers. This resulted in poor graph readers having to spend more time in the verification phase than good graph readers because they had to orient themselves to the graphs further in order to answer questions. Therefore, good graph readers were more successful in executing the processes involved in graph comprehension, thus allowing them to extract more complex information than poor graph readers in a shorter period of time.

Ratwani and Trafton (2008) empirically investigated the notion that similar graphs may share similar schemas using an undergraduate psychology student sample. They distinguished between the perceptual feature and invariant structure view. The perceptual feature view suggests that the pattern at the centre of a graphical display distinguishes one graph from another (e.g., bar charts have bars and line graphs have lines as the pictorial content) and so these differences in pattern will activate different schemas. The invariant structure view suggests that graphs that share similar frameworks will share the same schema. Bar and line graphs use the Cartesian coordinate framework and so will have the same schema, likewise pie and doughnut graphs share the same circle framework and so will have the same schema. Therefore, graphs based on different frameworks (e.g., line graph and pie charts) will not share the same schema.

They used a mixed cost paradigm where the measure for schema activation was time taken to answer a simple question when presented with either the same graph type in a block or different graph types in the same block. Ratwani and Trafton (2008) argued that if graphs are based on different schemas and a different graph type follows on from the previous type it will take time to load the appropriate schema, whereas if different graph types are based on the same schema there will be no time difference between these graph types. As there was no time difference between graphs that share structural similarities (i.e., the framework) but there was a time difference for graphs that did not share the same framework they concluded that graph schemas are based on structural similarities between graphs (invariant structure view).

Despite strong support for the major assumptions of Pinker's (1990) model it does have some weaknesses when applied to more complex semantically rich graphs. Firstly, although the model acknowledges that identifying and mentally keeping track of variable names is one of the processes involved in graph

comprehension, this stage is underemphasized, thought to occur at the final stage of graph comprehension and assumed not to influence comprehension beyond demands imposed on working memory. Furthermore, although the model accounts for differences between novices and experts and how knowledge will affect interpretation, the emphasis is on pattern recognition and whether this will activate knowledge of quantitative trends. There is no consideration of how content may affect interpretation when within context graphs are used, and how novices and experts will differ with familiar and unfamiliar content (Carpenter and Shah, 1998).

Carpenter and Shah's integrative model

Previous models of graph comprehension, what Carpenter and Shah (1998) refer to as the *pattern recognition models* emphasized the first and second stages involved in graph comprehension - encoding the visual pattern and inferring the quantitative trends the pattern is depicting – and placed minimal emphasis on the third process - identifying variable names and units plotted on axes and relating them to the lines at the centre of the display. According to these models these processes are performed sequentially and the most cognitively demanding step is the first process – pattern recognition.

Carpenter and Shah (1998) proposed the *integrative model* to counter the assumption that the most important stage in graph comprehension was the first pattern recognition stage. They criticize previous models emphasis on the pattern encoding stage and argued that the third stage involved in graph comprehension is just as important, if not more so than the early process of pattern recognition. Furthermore, they argued that rather than the processes involved in graph comprehension being executed sequentially comprehension is an integrative process requiring readers to repeatedly cycle between the pattern at the centre of the display and information plotted on axes. According to the integrative model the amount of cycles required depends on the number of distinctive visual chunks present in the graph.

Based on the Gestalt principles of perceptual organization Carpenter and Shah (1998) defined each line as a visual chunk. The model differentiates between similar visual chunks and distinct visual chunks. A distinct visual chunk is a line that is qualitatively different from other lines (i.e. lines with differing slopes). According to the integrative model lines with similar slopes (i.e. parallel lines) would only slightly increase time spent on interpretation whereas lines with dissimilar slopes (i.e. non-parallel lines) would increase time

significantly. The integrative model proposes that as graph complexity increases (i.e. complexity is defined by the number of distinct visual chunks presented in the graph) the number of cycles of pattern recognition, interpretation and integration also increase. According to the model each visual chunk would initiate a cycle of encoding, interpretation and integration. This process would be repeated for each chunk and the information would then be integrated together to form a coherent interpretation.

To test these assumptions Carpenter and Shah (1998) used three-variable line graphs with meaningful labels and units on the axes as their stimuli. The model makes a number of predictions. Firstly, to counter the assumption that the first pattern recognition stage is the main stage involved in graph comprehension participants' eye movements were recorded whilst they were interpreting graphs. The pattern recognition model would predict that the amount of time spent looking at variable names would be minimal, perhaps once at the end in order to identify variable names and range of values whereas the time spent looking at the pattern would take up the majority of time spent on interpretation. The integrative model predicts that although a large proportion of time will be spent looking at the pattern, the majority of time will be spent looking at variable names and scales on axes. Consistent with their hypothesis Carpenter and Shah (1998) found that a larger proportion of time was spent gazing at the title and x, y and z axes (i.e., the contextual information) than at the pattern.

Secondly, scan patterns supported their assumption that the processes involved in graph comprehension are integrative rather than sequential as participants' gazes moved repeatedly between the different regions and each major region was viewed multiple times. Furthermore, their prediction that dense graphs would take more time only if the lines were non-parallel was supported. They found that time spent on interpretation was significantly higher for dense graphs that depicted numerous non-parallel lines than dense graphs that depicted parallel lines.

A major strength of the integrative model is its emphasis on the final stage of graph comprehension - an under researched process. Using three-variable graphs with meaningful labels allowed Carpenter and Shah (1998) to investigate the effect of context on interpretation. Previous research up until this point used context free graphs (where axes were unlabelled or simply labelled using abstract letters). Furthermore, the model makes testable predictions which have been supported by empirical research and has provided a richer

understanding of the processes involved in comprehension using graphs that one would encounter in real world settings. However, the model does have some weaknesses. Carpenter and Shah (1998) focussed very narrowly on line graphs and so it would be reasonable to assume that some of the assumptions of the model will not apply to other graph formats. For example, the model emphasizes how in the case of three-variable line graphs quantitative information can automatically be retrieved for the variables plotted on the x-y axes whereas quantitative information for the variables plotted on the z-y axes are rarely spontaneously described and require inferential processes in order to interpret.

Peebles and Ali (2009) found that this effect was reversed when participants were required to interpret three-variable bar graphs. The z-y relationship was automatically retrieved but the x-y relationship was rarely described by participants. Therefore, in order to generalize to other graph types the model needs to account for how differences in visual properties affect which relationship a particular graph format makes salient. Finally, although the model highlighted the importance of the third process involved in graph comprehension, it does not adequately account for potential difficulties readers may experience when identifying referents. The model predicts that the majority of gazes will be on information plotted on axes and readers will repeatedly look at the same sections whilst interpreting each visual chunk and this assumption was supported using eye tracking. Similar to previous research the model predicts that the more variables readers have to encode the larger the burden on working memory, but does not go any further in predicting any potential difficulties readers may experience.

The experiment conducted by Peebles and Ali (2009) revealed that keeping track of referents is not the only problem readers experienced when attempting to interpret three-variable data sets. They found that novice readers are extremely poor at interpreting these types of graphs and a major problem readers experience in the line graph condition is relating the pattern at the centre of the display to variables plotted on axes. This resulted in participants completely ignoring one of the variables and focusing on two variables when interpreting three – variable graphs. Therefore, it would appear that novice readers struggle with the third process involved in graph comprehension – relating the pattern at the centre of the display to referents plotted on axes.

Although the integrative model demonstrated the importance of the third process involved in graph comprehension it does not predict potential problems novices will experience with this process when they are required to interpret line graphs. Furthermore, because of the limited focus on line graphs the integrative model does not distinguish between graph formats and how this may influence the third process involved in graph comprehension, thus limiting the scope and applicability of the model (Peebles and Ali, 2009).

Summary

The study of graph comprehension has naturally progressed in a linear fashion with early research attempting to isolate which graph format was universally preferable to another (Cleveland and McGill, 1984, Cleveland, 1985). These sweeping generalizations were shown to be inadequate however with further research demonstrating different perceptual tasks interact with graph format to determine judgment, speed and accuracy (Simkin and Hastie, 1987).

Further research focussed on the relationship between task requirements and how the display is organized with the proximity compatibility principle proposing there should be a close fit between mental processing and display proximity. This research provided guidelines concerning graph construction and how information should be organized in diagrams. These guidelines went beyond single factors such as graph format and considered the complex interaction of how task requirements interact with graph format to determine efficacy of various graphs.

Models of graph comprehension went further and attempted to explain how various factors interact to influence graph comprehension (Pinker, 1990, Carpenter and Shah, 1998) to provide an integrated theory of how each factor plays a role in reading and interpreting graphs. However, these models have either been limited in scope (Shah and Carpenter, 1995) or underemphasized crucial processes involved in graph comprehension (Pinker, 1990).

A review of the literature has revealed areas of interest that have received little attention in the graph comprehension literature. Traditionally, research in the field of graph comprehension has focussed on perceptual processes involved during graph reading and tasks have taken the form of simple fact retrieval in abstract or arbitrary domains. Very little attention has been paid to complex comprehension tasks

(Leindhardt, Zaslavsky & Stein 1990, Friel, Curcio and Bright, 2001). Lewandowsky and Behrens (1999) highlighted the need for research into comprehension of conceptual graphical tasks. In agreement with the current research question they suggest that interpretation of factorial experiments would shed light on conceptual graph understanding. The research literature abounds with factorial experiments whose results are typically displayed in line and bar graphs, yet the research into interpretation of factorial experimental results is extremely limited.

Therefore, the empirical work presented here aims to investigate conceptual understanding of complex graphs in an applied setting. The work presented here will move beyond past research to investigate how commonly used graphs are understood. These graphs will be meaningful in content and students who are expected to be able to use such diagrams will be employed as the sample to answer the research question. The aim is to ascertain how competent students are at interpreting these types of graphs and how misconceptions can be addressed to improve students' understanding of these types of displays.

Chapter 3 The verbal protocol method

Introduction

This chapter briefly outlines the different types of protocols which can be obtained when asking participants to think aloud. The theoretical assumptions underlying this approach are discussed in relation to how closely verbalization reflects thought processes and the key criticisms against these assumptions are reviewed. The advantages of employing the verbal protocol method as a process tracing method is outlined and discussed in relation to the research question being investigated.

The verbal protocol method

Arguably one of the most interesting developments since the cognitive revolution is the attempt to develop rigorous methodologies to trace thought sequences as a valid source of data on thinking. As well as being interested in the final product of thought (e.g., the decisions people make) cognitive psychologists are also interested in the processes which lead to the final thought output. This has led to an emphasis on developing methodologies which can trace the thought sequences leading to the output observed. Although verbal protocols were employed as a methodology to investigate cognitive processes involved in task completion prior to Ericsson and Simon's (1984) publication of "Verbal reports as data", the application of this methodology to tracing thought sequences has been attributed to their theory of protocol generation which provides extensive coverage of how different types of verbalizations elicit different types of responses.

Ericsson and Simon (1984, 1993) following on from Newell and Simon (1972) proposed the verbal protocol method as a means of tracing cognitive processes. The fundamental assumption underlying the verbal protocol method is that legitimate data on thought processes can be obtained when we instruct participants to verbalize their thoughts without significantly changing the sequence of thoughts involved in completing a task. Therefore, Ericsson and Simon (1984, 1993) make strong claims that the verbalizations participants provide – if instructions are followed carefully - are no different to the thought sequence that would have been followed if participants underwent the task silently. In their detailed theoretical account of protocol generation Ericsson and Simon (1984, 1993) distinguished between three different types of verbalizations to

predict which of these three would be appropriate to employ to trace cognitive processes and under which circumstances they would provide valid data.

Type 1 verbalizations are direct verbalizations. Participants simply report their thoughts and the information being reported is consistent with a verbal code. For example, a researcher may investigate the strategies employed in multiplication tasks. A participant may be told to multiply 108 by 9 and whilst calculating the answer they would be asked to report their thoughts. In this case a person may report "100 * 9 = 900 8 * 9 = 72 so the answer is 972". Intermediate steps may involve the participant going through the 8 or 9 times table. Ericsson and Simon (1984, 1993) argued that when Type 1 verbalizations are employed as a methodology where participants are asked to report their thoughts during a task then verbalizations will reflect the contents of short-term memory and there will be no change in the way information is heeded or reported due to the requirements to think aloud. This is because participants are only reporting thoughts that are being attended to and they are not required to describe or provide an explanation for the problem solving strategies being employed. In fact Ericsson and Simon (1984, 1993) warn against such demands on participants. Participants provide a protocol of what they are thinking whilst completing a task; it is the researcher's job to draw inferences about the processes involved in task completion.

Type 2 verbalizations require participants to recode information. The most often cited example is thinking aloud whilst solving an imagery task (e.g., the Raven Matrices). Manipulating such images and recoding thought sequences into a verbal code adds to the cognitive load of the task and can affect task completion. However, Ericsson and Simon (1993) argue that although such recoding may increase response times it does not alter the sequence of thoughts involved in task completion. In a review of numerous studies employing Type 2 verbalizations Ericsson and Simon (1993) found that the results were consistent with their predictions – the additional demands of recoding information increased response time to complete a task but the sequence of problem solving was not affected.

Type 3 verbalizations require participants to provide an explanation for their decision or the strategy they employed. It is this type of verbalization which can alter the sequence of thought processes involved in task completion as they require additional processing other than describing thinking whilst undergoing a task. These three types of verbalizations are employed in different ways depending on the research question being

investigated. The most common techniques require participants to think aloud concurrently throughout the task or retrospectively report their thoughts after task completion. Concurrent verbalizations are considered the most valid way of tracing thought processes as retrospective verbalizations are susceptible to forgetting or fabrication, particularly if the time lag between task completion and providing a protocol is too long. Therefore, Ericsson and Simon (1993) recommend whenever possible concurrent verbalizations should be used as they provide the closest reflection of thought processes mediating task completion.

According to Ericsson and Simon's (1984, 1993) theory of protocol generation, concurrent verbalizations, when conducted appropriately accurately reflect the processing which occurs throughout a task. In order to ensure this occurs participants must only report the thoughts that enter into attention whilst undergoing the task. If participants are asked to explain their thought processes supplementary information is drawn upon which changes how the task is performed because participants then need to think about information which is not normally accessed to complete a task. The concurrent verbalization method has been widely adopted in the research literature with thousands of research papers using this approach in an attempt to trace underlying processes involved in decision making, problem solving, text comprehension, diagrammatic reasoning, writing and various other areas (for a review see Crutcher, 1994).

In an analysis of how the verbal protocol method can be employed to uncover cognitive processes involved in writing Hayes and Flower (1983) distinguished between "process tracing" and "input-output" methods. Hayes and Flower (1983) use a metaphor of a locked room to demonstrate the problems with using input-output methods to investigate cognitive processes involved in writing. When these methods are employed we act as though the processes take place in a locked room – a room we cannot go into. We place the participant with the task and materials into the room (inputs) and we receive the finished product (the output) outside the room. In these cases researchers do not attempt to observe directly cognitive processes involved in the task. Instead, we rely on altering various inputs in an attempt to observe the effect on the output and then conjecture the processes involved in task completion.

However, when we use process tracing methods such as verbal protocols, as well as the information that we gain from input-output methods we can observe what is going on in the room and examine some of the

processes by which input leads to output. Hayes and Flower (1983) outline three key persuasive arguments for why we should employ process tracing methods.

First, process tracing methods give us a lot more information about processes than input-output methods. For example, we can identify the problems students may be experiencing if we ask them to verbalize their thoughts throughout a task than if we rely solely on analyzing the errors they made at the end of the task. Second, process tracing methods produce in-depth data which can provide excellent opportunities for exploring hypotheses in the early stages of the research process. During the process researchers may discover numerous findings or potential explanations for patterns which emerge in the data which they did not identify beforehand. Therefore, this method is considered especially useful when the aim is to generate hypotheses which would be difficult to predict a priori (Wilson, 1994). Third, there are some characteristics of processing which are problematic to examine without employing process tracing methods. For example, process tracing methods show us the order in which ideas are created during a task which can be very different from how these ideas are presented in for example, a written response. If only the output is analyzed it can be difficult to determine the order in which ideas emerged.

Despite the widespread use of this methodology in the literature the verbal protocol method has been criticized on a number of grounds. One of the most influential reviews and criticisms of the method was by Nisbett and Wilson (1977) whose paper "Telling more than we can know: Verbal reports on mental processes" reviewed studies which showed participants were unaware of why they made the choices they did or provided incorrect reasons for their behaviour – indicating they lacked access to their mental processes. They found that although participants could provide an explanation for their behaviour, often the explanation was inaccurate. Participants' behaviour was manipulated experimentally, then they were questioned about their behaviour and although participants provided valid reasons for their answer, few correctly attributed it to the experimental manipulation.

Nisbett and Schacter (1970, cited by Nisbett and Wilson, 1977) conducted one such experiment in which participants were asked to endure a succession of gradually escalating electric shocks. Before they received the shock one group was given a placebo pill which they were told would result in heart palpitations,

breathing irregularities, hand tremor and butterflies in the stomach (the most common symptoms associated with receiving an electric shock).

Nisbett and Schacter (1970) predicted that this group would believe the unpleasant symptoms they were experiencing was due to the pill and so would be more likely to endure a higher degree of shocks. The control group however, could only associate their unpleasant symptoms to the shocks they were receiving, so would be less likely to endure the shocks. This hypothesis was supported; those who had taken the placebo pill took four times the voltage of shocks compared to the no pill group. After they had completed the experiment participants in the placebo pill groups were interviewed. These were the typical questions they were asked and the responses they provided:

Question: "I notice that you took more shock than average. Why do you suppose you did?"

Typical answer: "Gee, I don't really know. Well, I used to build radios and stuff when I was 13 or 14 and maybe I got used to electric shock."

Question: "While you were taking the shock, did you think about the pill at all?"

Typical answer: "No, I was too worried about the shock."

Question: "Did it occur to you at all that the pill was causing some physical effects?"

Typical answer: "No, like I said, I was too busy worrying about the shock." (Nisbett and Wilson, 1977, p.237).

Nisbett and Schacter (1970) found that only 3 out of 12 participants realised they may have endured more shocks because of the symptoms they believed the pill would produce. The experimenter then provided participants with details of the experimental hypothesis telling them that they (researcher) believed they (participant) would attribute negative symptoms to the placebo pill. The experimenter then asked the participant if they had thought what he had told them. Participants generally stated that although some people would go through such a process, they had not. Nisbett and Wilson (1977) cite numerous studies which follow this pattern – participants provide reasons for their answers but cannot correctly infer what is influencing their behaviour. Based on their review they concluded that participants had no greater access to

their mental processes than anyone else – thus calling into question the use of introspective methods to describe cognitive processes.

Ericsson and Simon (1993) challenged Nisbett and Wilson's (1977) review by arguing that their findings are a result of ineffective procedures for obtaining verbal reports. For example, in a number of studies cited by Nisbett and Wilson (1977) the answer to questions could be produced without the participant needing to consult cognitive processes involved in performing the task. As an alternative to reflecting on memory participants can draw on background information to answer the question. In the example of the experiment requiring participants to endure shocks, participants were asked "I notice that you took more shock than average. Why do you suppose you did?" (Nisbett and Wilson, 1977, p.237). Ericsson and Simon (1993) argue that it is not clear to them (and so the participants in the study) that memory of the cognitive processes should be the source they should draw upon for the answer to the question. Ericsson and Simon (1993) propose that if participants can provide an answer to questions without interrogating their memory for why they did what they did, they may prefer this option to retrieving their answers from memory.

Furthermore, the procedures employed for eliciting retrospective protocols from participants in studies reviewed by Nisbett and Wilson (1977) violated recommendations for how retrospective protocols should be used due to the rapid decay of memory traces. In their theoretical framework proposing how verbal protocols reflect thought processes Ericsson and Simon (1993) proposed that participants report information from short – term memory if concurrent verbal protocols are employed or if retrospective protocols are employed immediately after task completion.

In the majority of studies reviewed by Nisbett and Wilson (1977) however, the time interval between task completion and request for verbal reports meant that plausibly the information was no longer available in short-term memory. Therefore, the inclination to provide verbal reports by tapping into memory processes will be less likely to occur when the memory trace for information needed is weak. If the experimenter does not provide specific probes cueing the relevant aspects of memory participants are likely to rely on background knowledge (if available to them) to provide the response.

Finally, a lot of the research reviewed by Nisbett and Wilson (1977) required participants to describe information that cannot be provided even when participants have full access to mental processes – that is participants are asked "why" questions concerning causes for their behaviour. Type 1 and Type 2 verbalizations prohibit encouraging participants to infer causes for their behaviour. It is only for Type 3 verbalizations – where participants are asked to explain or justify their actions (e.g., why do you prefer this painting over another?) that the problems outlined by Nisbett and Wilson (1977) could potentially emerge.

When the research employs Type 1 or type 2 verbalizations, either concurrently or retrospectively, participants simply report their thoughts without explaining or justifying what they are doing. Ericsson and Simon (1993) argue that the effects of Type 3 verbalizations should not be generalized to Type 1 and 2 verbalizations. For this research participants will be required to think aloud concurrently throughout the task without explaining why they used the strategy they did (Type 1 verbalization). Therefore, the criticism that participants are unable to correctly identify stimuli influencing their behaviour is not an issue for this research question.

A second criticism against this method concerns completeness of protocols. If some processes occur unconsciously, then they will not be available to the participant to report. For example, we can recognize a person but most likely not be able to report how this recognition occurred. This is especially an issue for research investigating expertise. Processes may become so automated that participants may not need to attend to certain information (e.g., individuals learning to drive will be consciously focusing on the actions they need to take whereas for an experienced driver some of the moves will be automated).

Ericsson and Simon (1993) accept that some information is not available to report as verbal protocols can only reflect information which reaches consciousness. Information which never reaches consciousness cannot be accessed for verbal reports. Therefore, this method is not appropriate to study certain phenomena (e.g., learning without awareness). In rebuttal of the criticism that verbal protocols are incomplete, Hayes and Flower (1983) question this line of argument. Although they accept that protocols may not be complete they argue that it is paradoxical that this method is singled out for this particular criticism because protocols tend to provide more data than the methods they are compared to. Therefore, if one were to adopt a

comparative approach then the same criticism could be leveled against other potential methodologies adopted to investigate the question of interest.

Although there is a dispute between Ericsson and Simon's (1993) and other researchers' stances (e.g., Wilson, 1994) regarding just how much of a problem unconscious processing is, this particular criticism does not affect the question this research is investigating. Graph interpretation is not a highly automated skill for novices. The participants will have to attend to the information they are viewing in order to complete the task. Therefore, information will be available in short-term memory for participants to report as they undergo the task.

The final major criticism concerning the verbal protocol method - in all its forms - is the idea that providing a protocol can be "reactive". Reactivity effects are when thought sequences involved in undertaking a task are altered due to the demands of thinking aloud. Ericsson and Simon (1993) vigorously deny Type 1 and Type 2 verbalizations are susceptible to reactivity effects, arguing that the requirements to think aloud will not alter sequence of thought processes. They cite numerous studies (e.g., Norris, 1990, Biggs, Rosman and Sergenian, 1993, Sanderson, 1990) demonstrating that employing the concurrent think aloud method reveal no effect on cognitive processes (established by comparing the performance of think aloud participants to those undertaking the task in silent conditions). However, a handful of studies have revealed that the requirement to verbalize during a task can affect cognitive processes (Wilson, 1994). Some researchers have argued that requiring participants to verbalize can direct individuals' attention to information that is easily accessible and also easy to verbalize.

For example, Schooler, Ohlsson and Brooks (1993) investigated how providing retrospective and concurrent protocols whilst solving insight problems would affect performance. They compared a group of participants providing retrospective protocols with a silent control group. Participants stopped midway through the problem solving tasks and were required to provide a retrospective protocol of how they had been trying to solve the problem. A silent control group was also distracted midway through the task and was required to engage in an unrelated activity. The findings showed that those who were required to provide a retrospective protocol performed worse (solving fewer problems) than the silent control condition on insight problems but there was no difference for non insight problems.

Schooler, Ohlsson and Brooks (1993) argued that "verbal overshadowing" occurs when participants are required to solve insight problems. Verbal overshadowing is when attention is directed to information that can easily be verbalized and so eclipses information that cannot easily be put into words. Schooler, Ohlsson and Brooks (1993) propose that insight problems involve a number of elements not amenable to verbal reporting. This then directs participants' attention to information which is reportable, but not helpful for producing solutions to insight problems. Their findings led them to conclude that crucial aspects of problem solving are "overshadowed" by the demands of verbalization. They argue that thinking aloud - either concurrently or retrospectively – interferes with difficult to verbalize processes that are crucial for successful completion of insight problems.

Ericsson and Simon (1993) countered this explanation with an alternative explanation. They argue that insight problems involve reaching the solution suddenly – individuals need to overcome an erroneous first assumption of how to solve the problem by recalling new information. This indicates that at the time participants were required to provide a retrospective protocol in the tasks set by Schooler, Ohlsson and Brooks (1993) participants most likely would have only retrieved information that is applicable to inaccurate strategies. Therefore, the act of verbalizing these inaccurate strategies would have reinforced them. When participants continued solving the problem they would be at a distinct disadvantage because the reinforced incorrect approach would make it more difficult for them to retrieve the new information relevant to solving the problem. Ericsson and Simon (1993) argue that if retrospective reports had been utilized the way they recommend – after task completion, these problems could have been avoided.

However, a further experiment conducted by Schooler, Ohlsson and Brooks (1993) did not suffer from these problems. They required participants to think aloud concurrently throughout the task – rather than interrupting them and asking them for a retrospective report. They found that participants who provided a concurrent protocol did not differ from the silent condition in performance when completing non-insight problems. However, those who verbalized were 25% less likely to reach a correct solution for insight problems compared to the silent group. Ericsson and Simon (1993) struggled to explain these results, suggesting the effect needed to be replicated. They argued it was possible that a minor deviation in the instructions given to participants could potentially explain the effect. Participants were asked to think aloud

including any information they read and any questions they asked themselves. Ericsson and Simon (1993) suggested that reading aloud the instructions could have potentially interfered with the retrieval of new information required to solve insight problems.

Another study requiring participants to think aloud concurrently whilst completing a range of tasks was conducted by Russo and Johnson (1989) who empirically tested the assumptions Ericsson and Simon (1984) outlined in their theory of protocol generation. They found that Ericsson and Simon's (1984) theoretical assumptions of when protocol generation would be reactive failed to accurately predict the reactivity observed in the tasks they employed. Russo and Johnson (1989) found an interaction between task type and accuracy. Interestingly, one of the two tasks which were found to be reactive should not have been according to Ericsson and Simon's (1993) analysis of protocol generation. They found that providing a concurrent protocol significantly improved the accuracy of a choice between two gambles but conversely decreased the accuracy of adding three-digit numbers. The other two tasks showed no reactivity effects.

Ericsson and Simon (1993) admitted these results were "puzzling". However, they drew attention to the fact that Russo and Johnson (1989) required participants to talk continuously so when they fell silent they prompted them after a few seconds. Ericsson and Simon (1993) state that verbalization should be secondary to completing the task so researchers should wait 10-15 seconds before requesting participants to keep talking so any interference does not occur. Although Russo and Johnson (1989) only needed to prompt participants a few times Ericsson and Simon (1993) argued that participants could potentially have altered their cognitive processes during practice trials to be able to talk continuously. Again Ericsson and Simon (1993) suggested replication was necessary as numerous studies reviewed have shown no reactivity effects of addition problems.

To summarise, the verbal protocol method has been widely adopted in the research literature as it can provide researchers with the means to trace cognitive processes involved in various types of tasks. Although the verbal protocol method has been criticized on a number of grounds these criticisms have been addressed in Ericsson and Simon's (1993) review of the research literature. Importantly, some key criticisms are not applicable to this research question so do not affect the use of this methodology. There are a few studies where reactivity effects have emerged however (e.g., Russo and Johnson, 1989, Schooler, Ohlsson and

Brooks, 1993) which indicate further research is required into whether this methodology is appropriate for certain tasks. However, this is not the case for the tasks used in this research as studies in diagrammatic reasoning have frequently made use of the verbal protocol method (e.g., Shah and Carpenter, 1995, Shah, Mayer and Hegarty, 1999, Trafton et al, 2000, Peebles and Ali, 2009).

The think aloud method was employed for this research question because it allows the tracing of processes which can yield rich data and allows us to trace important cognitive processes which would be difficult to observe using other measures (Flower and Hayes, 1981, Ericsson and Simon, 1993, Crutcher, 1994). Specifically, this methodology was used to investigate how differences in bar and line representations influence the interpretation provided by novice readers of the relationships depicted in these diagrams. Using a process tracing method will allow an analysis of which features of the representation result in the errors observed in the Peebles and Ali (2009) study which I predict will again emerge in the experiments reported here. Although other methods (e.g., question answer tasks, drawing tasks) would have allowed me to record students' interpretations of these graphs, I would not have been able to trace the underlying cognitive processes leading to the errors students make whilst attempting to provide an interpretation (Crutcher, 1994, Payne, 1994).

Chapter 4

Students' understanding of interaction graphs

Factorial research designs are widely used in all branches of the natural and social sciences as well as in engineering, business and medical research. The efficiency and power of such designs to reveal the effects and interactions of multiple independent variables (IVs) or factors on a dependent variable (DV) has made them an invaluable research tool and, as a consequence, the teaching of such designs, their statistical analysis and interpretation lies at the core of all natural and social science curricula.

The simplest form of factorial design is the two-way factorial design, containing two factors, each with two levels, and one DV, for example the differences in word recall (DV) between amnesics and a control group (IV1) in an implicit versus explicit memory task (IV2). Statistical analysis of these designs most often results in a 2x2 matrix of mean values of the DV corresponding to the pairwise combination of the two levels of each IV. Interpreting the results of even these simplest of designs accurately and thoroughly is often not straightforward however, but requires a significant amount of conceptual understanding - for example, the concepts of simple, main, and interaction effects. As with most other statistical analyses however, interpretation can be eased considerably by representing the data in diagrammatic form.

Data from two-way factorial designs are most often presented as either line or bar graphs, variously called interaction or ANOVA graphs. Examples of such bar and line graphs (taken from the experiments reported here) are shown in Figure 9. Bar and line graphs such as those in Figure 9 can display the same data set in the same coordinate system and are informationally equivalent (Larkin & Simon, 1987). In terms of their visual and conceptual structure, bar and line graphs have a great deal in common, the key difference being the way in which the data points are represented in the coordinate system. However this relatively minor difference has been shown to have a remarkable effect on which features are made salient, which in turn influences the type of information extracted from the display.

In line graphs, lines integrate individual plotted points into single objects, features of which (e.g., slope, height relative to other lines, etc.) can indicate relevant information about the entire data set (Carswell &

Wickens, 1990, 1996). This feature has been found to lead people to encode the lines in terms of their slope (e.g., Simcox, 1983, reported by Pinker, 1990) and interpret them as representing continuous changes on an ordinal or interval scale (Zacks & Tversky, 1999, Kosslyn, 2006). For this reason line graphs are typically regarded as a form of configural or object display. In contrast, bar graphs are an example of a separable display as each data point is represented by a single, separate bar. Because of this, people are more likely to encode bars in terms of their height and interpret them as representing the separate values of nominal scale data (Culbertson & Powers, 1959; Zacks & Tversky, 1999).

These differences in encoding and interpretation can result in significant performance variation for different tasks; people are typically better at comparing and evaluating specific quantities using bar graphs (Culbertson & Powers, 1959; Zacks & Tversky, 1999) whereas people are generally better at identifying trends and integrating data using line graphs (Schutz,1961).

This situation is therefore a prime, real-world example in which two informationally equivalent and relatively similar representations are widely used, but which is known to be computationally inequivalent (Larkin & Simon, 1987) in certain circumstances. It seems appropriate to ask therefore, whether these computational differences significantly affect the ease and efficiency with which people interpret them and the depth and accuracy of the interpretations produced.

According to the proximity compatibility principle (Carswell & Wickens, 1987), graph format should correspond to task requirements, so that configural displays should be used if information needs to be integrated, whereas separable displays are more appropriate if specific information needs to be located. In the case of interaction data however, there are reasonable arguments for using either format.

Interaction graphs differ from more conventional line graphs in that the variables plotted on the x axis are categorical, regardless of whether the underlying scale could be considered as continuous (e.g., hot/cold) or categorical (e.g., male/female). The argument for using bars for interaction graphs is that, because people encode bars as separate entities, they are less likely to misinterpret the levels of the x axis variable as representing two ends of a continuous scale (Zacks & Tversky, 1999, Aron, Aron, & Coups, 2006). By contrast, line graphs are more likely to be interpreted as representing continuous data with points on the lines

representing intermediate values on the scale. Proponents of the line graph (e.g., Kosslyn, 2006) have argued however that the risk and costs of misinterpreting line graphs are outweighed by the benefit of lines for producing easily recognizable patterns that can be associated with particular effects or interactions.

A reading of the academic psychology research literature suggests that bar and line interaction graphs are used roughly equally. To test this impression, I counted the number of bar and line interaction graphs in the 2009 volumes of two journals widely recommended to undergraduate students as academic sources and which together cover a broad range of topics and methodological practices; *Psychological Science* and the *Journal of Experimental Social Psychology*.

The analysis revealed that this was generally the case. The mean number of interaction bar and line graphs per issue of Psychological Science were 11.83 (SD = 5.89) and 16.83 (SD = 5.27) respectively while those for the Journal of Experimental Social Psychology were 25.17 (SD = 11.75) and 24 (SD = 24.40) respectively. Taking the two journals together, the proportions reveal a slight preference for line graphs (54%) over bar graphs (46%).

This preference was found to be more pronounced in undergraduate psychology textbooks however. A similar analysis carried out on two current popular psychology textbooks used in the undergraduate Introduction to Cognitive and Developmental Psychology class at the University of Huddersfield (Eysenck & Keane, 2005, Boyd & Bee, 2006) found that line graphs were favoured 20% more than bar graphs.

Which diagram to use for displaying two-way factorial design data may not always be down to an explicit rational decision by the user but may often be constrained by external factors. For example, one of the most popular statistical software packages in academic use, PASW Statistics (produced by SPSS inc.) provides only the line graph option as part of its ANOVA functions. It is not unreasonable therefore, to assume that undergraduate students are more likely to be required to use the line graph format when analyzing their own data and to comprehend them in some detail in order to interpret their experimental results.

If the visual properties of line graphs can lead users to focus on features that suggest incorrect interpretation (e.g., a continuous valued x variable) or distract attention away from the plotted data points, then they may

not be the best representation to use, particularly in educational settings where novice users are learning how to analyze and interpret the various relationships.

When attempting to compare and evaluate performance with different graphical formats, it is essential to have a set of behavioural criteria or categories with which to do so. From the considerable number of studies conducted into graph comprehension a consensus has emerged on the broad three-level taxonomy of skills required for the task. In a review of five studies, Friel, Curcio, and Bright (2001) characterized the three levels as elementary, intermediate, and advanced (or more descriptively as “read the data", “read between the data" and “read beyond the data" respectively).

At an elementary level people focus primarily on extracting specific values. At an intermediate level people interpret the data presented more fully and, to a certain extent at least, integrate the information together. At an advanced level people also make inferences beyond what is explicitly stated in the graph by hypothesizing based on trends depicted in the graph.

While there will always be differences between individuals in terms of their general graph sense (Friel et al., 2001), a characteristic that develops with experience over time and involves knowledge of such things as how coordinate systems work and general rules of labelling by colour etc., it is reasonable to assume that individuals will differ in terms of their ability to interact with different graph types. This can be for a number of reasons; familiarity, particular idiosyncrasies of the representation, or the structure of the data being presented. For example, if individuals are unfamiliar with the particular representational features of a format, then they may only be able to interact at an elementary level with the only option available being to read off individual values.

Experience of teaching undergraduate psychology students to interpret two-way factorial data with the line graphs found in common statistical software provides at least anecdotal evidence that this is indeed the case. I have typically found that students who have little difficulty working at an intermediate level with line graphs when they represent continuous or interval data, may only be able to produce elementary performance with two-way factorial line graphs. Furthermore, it seems that this discrepancy in performance can persist despite substantial amounts of exposure, with many students continuing to have difficulty interpreting the

line graphs accurately and often only being able to obtain a superficial and incomplete understanding of the relationships between the variables.

For example, I have observed that students will often be able to identify and reason about the variable represented in the legend (e.g., the Stimulus Type variable in Figure 9a) but fail to do so for the variable represented on the x axis (the Task variable in Figure 9a). One explanation for this is that the plot lines distract attention away from the more relevant graphical features (the points at the ends of the lines) and then to the value labels in the legend rather than to the labels under the points on the x axis (Peebles & Ali, 2009).

There is reason to believe that this pattern of interaction may not be found with bar graphs however. Peebles (2008) recently demonstrated using a mixed sample of staff and students that people perceive informationally equivalent bar and line graphs quite differently. For example, when required to compare values plotted in bar and line graphs with an average (represented as a line drawn from the y axis parallel to the x axis), bar graph users significantly underestimated the size of the plotted value relative to the mean compared to line graph users. The effect occurred despite the fact that the values being compared were plotted at exactly the same locations in the two graphs and was explained as resulting from a process whereby bar graph users' visual attention was drawn to the length of the bars as they extend from the x axis (cf. Pinker, 1990; Simcox, 1983) rather than to the distance between the top of the bar and the mean line - thereby accentuating the perceived difference between them.

The fact that the bars in bar graphs are attached to the x axis may provide a more balanced representation in which the graphical features index both IVs more evenly. To test this hypothesis Peebles and Ali (2009) conducted an experiment in which people were asked to interpret informationally equivalent bar or line graphs representing two-way factorial design data as fully as possible while thinking aloud. Analysis of the verbal protocols revealed significant differences in how people interpreted the two graph formats. Specifically, it was found that 39% of line graph users were either unable to interpret the graphs, or misinterpreted information presented in them. No bar graph users performed at this level. This finding led them to propose a fourth, lower category of comprehension ability which was termed "*pre-elementary"*.

The main error produced by the pre-elementary line graph users was what I had noticed anecdotally in statistics classes - ignoring the x axis variable entirely or ignoring one level of the x axis variable. Additionally, they found that bar and line graph users identified different IVs as the primary focus of their interpretation; line graph users typically used the legend variable whereas bar graph users were more likely to use the x axis variable.

Peebles and Ali (2009) argued that this reversal effect is due to different Gestalt principles of perceptual organization acting in each graph. In the case of bar graphs, the x variable values are grouped together on the x axis and, by the Gestalt principle of proximity (Wertheimer, 1938) each cluster of bars forms a separate visual chunk. Participants identify these chunks, access the associated label and then use them as the values by which to compare levels of the z variable (e.g., in xix (b) a user may say “with hot temperature, high stress produces a lot more fractures than low stress").

In the case of line graphs however, data points are connected by the lines which, by the Gestalt principle of connectedness (Palmer & Rock, 1994), form individual visual chunks. This leads users to identify rapidly these chunks, access the associated label in the legend by colour and then use them as the values by which to compare levels of the x variable (e.g., in Figure 9a a user may say “with word stimuli, response time is much faster in task AA than for task AB").

I have taken these findings as providing preliminary evidence that the representational features of bar and line interaction graphs strongly influence their interpretation and result in marked differences in people's ability to comprehend the relationships depicted fully and accurately. In addition, these results suggest that the two graph formats produce significantly different patterns of interaction, with users' attention being attracted to different variables and regions of the graph.

Experiment 1

Experiment 1 is a replication of the experiment conducted by Peebles and Ali (2009). Although providing valuable initial insights, the experiment had one main limitation; the 29 participants were drawn from both staff and students from the University of Huddersfield with a wide age range (23.1 to 62.2, M = 42.8, SD = 12.7), with a majority (48.3%) being academic staff from different schools in the university with smaller proportions of non-academic staff (20.7%), postgraduate (20.7%) and undergraduate (10.3%) students. Therefore the sample varied widely in terms of their exposure to data analysis in general and interaction graphs in particular from complete novices to experts.

As the primary aim of this research is to determine how graphical features affect relatively novice users - particularly in an educational context - a more homogeneous sample taken from an appropriate student population will provide a more accurate indication of the proportion of students that cannot understand these types of graphs accurately. It will also allow a more precise measure of the specific effects of graph format on comprehension by minimizing the potentially confounding effects of familiarity and expertise. The focus will be on errors participants make rather than the different statement patterns bar and line graph users produce.

The aim of the first experiment therefore is to compare the levels and patterns of comprehension between undergraduate psychology students using informationally equivalent three-variable bar and line interaction graphs. This will not only assess the robustness and generalisability of the initial findings reported in Peebles and Ali (2009) but also determine whether the differences are more pronounced in undergraduate students. Experiment 1 therefore is a replication of the Peebles and Ali (2009) experiment using an undergraduate psychology student sample.

Method

Participants

Forty-two undergraduate psychology students (36 female, 6 male) from the University of Huddersfield were paid £5 in supermarket vouchers to take part in the experiment. The age of participants ranged from 18.8 to 37.1 years with a mean of 21.2 years (SD = 3.8). 21 participants were in their first year and 21 were in their second year of a three-year psychology degree.

Design

The experiment was an independent groups design with two between-subject variables: type of diagram used (bar or line graph) and the allocation of independent variables to the x axis and legend (labelled `normal' and `reversed'). Twenty-one participants were alternately allocated to each of the two graph conditions by alternating which graph condition each participant saw. For example, if participant one was in the line graph condition participant two was in the bar graph condition and so on. There were 11 participants in the normal-bar condition, 11 in the normal-line condition, 10 in the reversed-bar condition and 10 in the reversed-line condition.

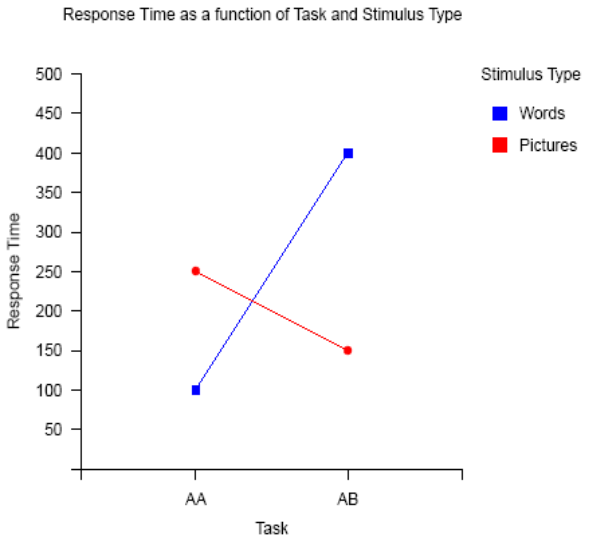
Materials

The experiment was carried out using a PC computer with a 43 cm display. The stimuli were twelve bar and twelve line three-variable interaction graphs depicting a wide range of (fictional) content. The graphs were approximately 18.5cm cm wide by 16 cm high and were drawn black on a light grey background with the legend variable levels coloured red and blue.

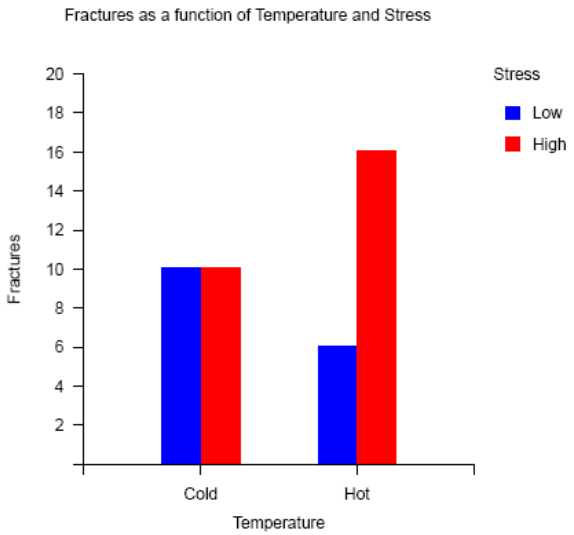
The variables and levels of each data set are shown in Figure 9. The numerical values for the variables were selected in order to provide the range of effects, interactions and other relationships between three variables commonly encountered in these designs (typically depicted in line graphs as parallel, crossed and converging lines, one horizontal line and one sloped line, two lines sloping at different angles, etc.)

The six normal bar and line graphs had IV1 on the x axis and IV2 in the legend whereas the six reversed graphs had the reverse allocation. This counterbalancing was undertaken as a precaution against the

possibility of any particular variable being more readily interpreted as continuous or interval data, thereby possibly biasing interpretation of the line graphs. Stimuli were presented by a computer program and participants' verbal protocols were recorded using the computer's digital audio recorder.

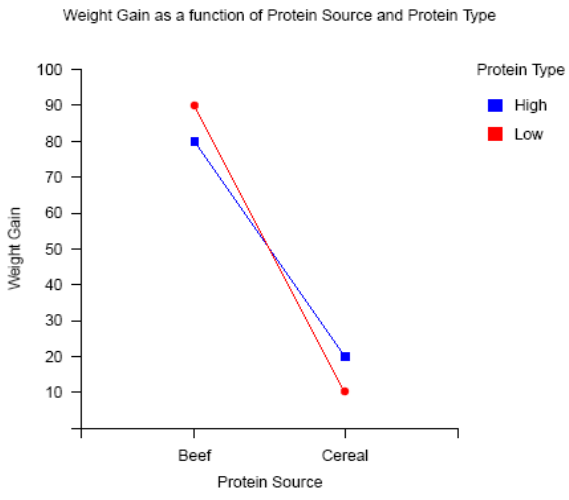


(A)



(B)

(C)



(D)

Figure 9: Bar and line graphs representing four of the six data sets used in Experiment 1. All graphs are in the 'normal' orientation

Procedure

Participants were informed that they were to be presented with a sequence of six three-variable line graphs and that their task was to try to understand each one as fully a possible while thinking aloud. The nature of the task was further clarified by telling participants that they were being asked to try to understand the relationships between the variables (rather than simply describing the variables in the graph), to try to comprehend as many relationships as possible, and to verbalize their thoughts and ideas as they did so.

Participants were instructed that when they felt they had understood the graph as much as possible, they should try to summarise the graph in just one or two sentences before proceeding to the next graph. They were also requested not to just skip a trial if they felt that they did not fully understand the graph but to try to interpret as much of it as they could.

During the experiment, if participants went quiet, the experimenter encouraged them to keep talking or asked them what they were thinking. If participants stated that they could not understand the graph, it was suggested that they attempt to interpret the parts of the graph they could understand. In addition, if participants' verbalizations consisted solely of descriptions of visual features or variable names, the experimenter encouraged them to try to understand the relationships between the variables. If they still could not do this, they were allowed to move on to the next trial. When participants had understood the graph as much as they could, they proceeded to the next trial by clicking the mouse on the graph. The graphs were presented in random order.

Results

Pre-elementary performance

Participants' protocols were transcribed and their content analyzed. Only statements in which a sufficient number of concepts could be identified were included for analysis. For example, the statement “Wellbeing is higher for high exercise than low exercise" was included whereas “Wellbeing is higher when. . . um. . . I'm not sure" was not.

If participants’ statements were correct but the interpretation only described part of the graph then the trial was scored as a correct interpretation providing all three variables were taken into consideration and a variable was not ignored. For example, the statement “for words task Aa produces a faster response time than Ab” (figure 9a) only describes one set of relationships in the graph as the relationship between stimulus type and task Ab is not described. These trials were scored as being a correct interpretation because participants demonstrated they were able to incorporate all three variables into their interpretation.

Similar to the above example, statements describing maximum or minimum values were also scored as correct interpretations. For example, the statement “Low protein beef results in the highest weight gain” was classified as a correct interpretation. Again, this was because all three variables were taken into consideration. If participants located minimum or maximum values but ignored a variable the trial was scored as incorrect. For example the statement “high protein type results in lower weight gain” was classified as an incorrect interpretation because the protein source variable is being ignored (refer to Item 1 in the appendix for an example of a scored transcript from the line graph condition).

Similarly, if participants were incorporating all three variables into their interpretation but were incorrect in their interpretation of the direction of effect (for e.g., if they stated increasing when the variable was decreasing) the trial was coded as a correct interpretation. This misinterpretation was only observed for the graph depicted in figure 9a and is consistent with findings in the literature that higher is better (the graph depicted in figure 9a depicts a slower response time further up the y axis and so presumably participants assumed the higher the better, i.e., quicker response time), Tversky, (2001).

Data analysis was conducted according to the procedure and criteria employed in the original Peebles and Ali (2009) study. For each trial, the participant's statements were analyzed against the state of affairs represented by the graph. If a participant made a series of incorrect statements that were not subsequently corrected, then the trial was classified as an `incorrect interpretation'. If the participant's statements were all true of the graph or if an incorrect interpretation was followed by a correct one however, then the trial was classified as a `correct interpretation'. In this way, each participant's trials were coded as either being correctly or incorrectly interpreted.

The verbal protocol for each trial was initially scored as being either a correct or incorrect interpretation by the author and a sample (approximately 25% from each graph type) was independently scored by another researcher. The level of agreement between the two coders was 95.3%. A Cohen's kappa test was conducted and revealed strong inter-rater reliability agreement between coders ($k = 0.90$, $p < .001$). When disagreements were found the raters came to a consensus as to the correct code.

This measure was then used as the basis for subsequent categorization into elementary and pre-elementary groups. For the purpose of the analysis, I classified participants as pre-elementary for their graph type if they interpreted 50% or more trials incorrectly (i.e., at least three of the graphs were classified as incorrect interpretations). This criterion was considered appropriate because it indicates that the user is unable to produce an accurate description of the data (even such basic information as point values) after at least two previous encounters with the same graph type - suggesting a lack of understanding of the basic representational features of the format (rather than just the content of the graph) and resulting in comprehension performance that does not meet elementary level criteria (Friel et al., 2001).

According to this classification criterion, 62% of the line graph users were pre-elementary compared to 24% in the bar graph condition.

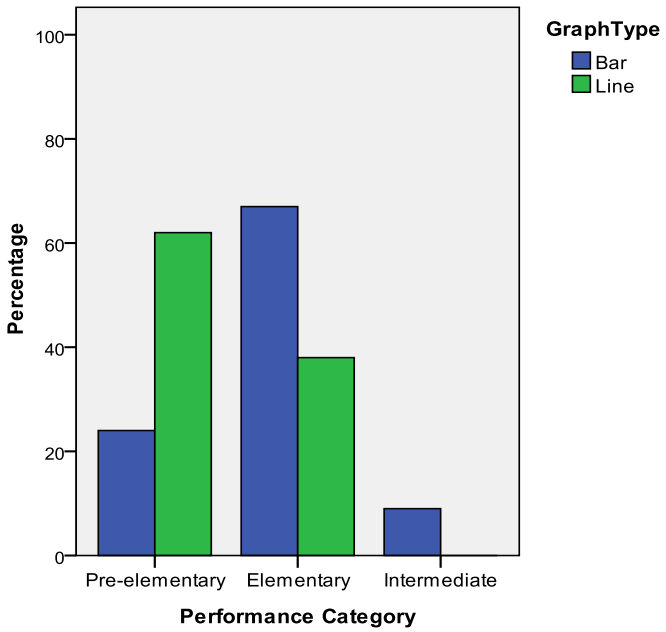


Figure 10 Percentage of bar and line graph users in the three performance categories, Experiment 1

A chi-squared test of independence[1] (one-tailed) revealed that this association between line graph users and pre-elementary performance was statistically significant (chi-square = 6.2; df = 1; $p < .05$), replicating the result of the original Peebles and Ali (2009) experiment. There was no significant association between performance and year of study (chi-square test of independence = 1.17; df = 1; $p = .20$), nor between whether they saw `normal' or `reversed' graphs (Fisher's Exact Test, $p = .27$ (line) and $p = .26$ (bar)).

To determine that these differences were not simply an artefact of classification of participants into pre-elementary and elementary categories, I also compared the number of correct trials between the two graph conditions. Non-parametric tests were used as data violated assumptions of normality and equality of variance.

A Mann Whitney U test (one-tailed) revealed that the number of correctly interpreted trials in the bar graph condition (mean ranks = 25.26) was significantly greater than in the line graph condition (mean ranks = 17.74), $U = 141.5$, $p < .05$.

[1] Chi-square conducted on the number of trials a particular error was made not the number of participants making a particular error.

In addition to this trial-level performance analysis, I also analyzed the nature of the errors made in incorrectly interpreted trials. When participants made an erroneous interpretation that was not subsequently corrected, in addition to the trial being classified as an `incorrect interpretation', the type of error was coded against the trial. The nature of the fault was categorized according to which of the variables had been ignored or misrepresented or whatever other error had occurred (see Table 2). Errors followed a similar pattern to the original experiment. Below I describe each error type, providing example statements and suggesting explanations.

Error analysis

Ignoring the x variable

Consistent with the original findings of the Peebles and Ali (2009) study a substantial proportion of line graph users (16.7%) described the effect of the legend variable and ignored the x axis variable altogether. This was the most common single error in the line graph condition, made by twice as many line graph users as bar graph users.

An example of this type of error for the line graph in Figure 9a is “Response time for words is increasing whereas for pictures it's decreasing". This statement simply describes the slopes of the blue and red lines respectively as read from left to right and does not explicitly identify any information regarding the levels of variable on the x axis.

Ignoring the z variable

This error can be considered the opposite of the previous one and occurs when participants describe the effect of the x axis variable but ignore the legend variable. An example of this type of error for the graph in Figure 9a is “Response time for task AA is increasing whereas for task AB it is decreasing". As with the previous error, the user is simply describing the slopes of the lines, but in this case associating each line with a level of the x variable. Compared to the corresponding x variable error, the proportion of participants producing this error was approximately equal between the two graph conditions, with the number of line graph users doing so dropping by roughly 50%.

Although ignoring one of the IVs will always produce an erroneous interpretation, depending on the data, some statements may be limited while also being a true description of the graph. For example, the statement “beef causes a higher weight gain than cereal" for Figure 9d is correct. However, if it was produced without any further elaboration or qualification, the interpretation is limited because the effect of both protein source and protein type have not been taken into account, and ignoring the effect of the latter on weight gain results in the interpretation being incomplete.

Content-specific errors

Two of the graphs resulted in specific patterns of error that are interpreted as being related to the nature of their content. The first concerns the relationship between temperature, stress and fractures (Figure 9b). I observed a number of participants producing statements indicating that they thought that the two IVs were causally related (i.e., temperature increasing stress) and omitting the dependent variable (fractures). An example of a typical statement was a participant saying “As temperature increases, so does stress, whereas cold doesn't affect stress".

Table2. Percentage of erroneous and missed trials for line and bar graphs, Experiment 1.

	Graph Type	
Error	Line	Bar
Ignoring the x variable	17.46	7.14
Ignoring the z variable	8.73	7.94
Content-specific errors	8.73	8.73
Miscellaneous errors	3.97	3.97
Missed trials	9.52	3.17

The second instance occurred for the graphs depicting the relationship between protein type, protein source and weight gain (Figure 9d). In this situation, a number of participants combined the variables plotted on the x axis and the legend because they assumed that high protein was associated with beef (protein source) but

associated low protein with cereal. In these trials, participants usually said something along the lines of “beef is a high protein type and so causes a higher weight gain, whereas cereal is a low protein type and so results in a lower weight gain".

In both cases, these errors can be explained as resulting from participants' prior knowledge of the variables and their possible causal links - in the former, the connection between temperature and stress in some materials and the latter that beef is a relatively high source of protein. However, in both instances, the number of these errors was low and was even between graph conditions (8.7% for both errors in both the bar and line graph condition). In addition, the number of errors unrelated to content for these two graphs far outweighed these content-related errors.

Miscellaneous single errors

An error was categorized as `miscellaneous' if participants were relating all three variables to each other, but their interpretation was incorrect, either because they were relating the variables incorrectly, or because their description was not consistent with the information in the graph. Miscellaneous errors, unlike the previous errors, were not systematic in that each error categorized as being miscellaneous only occurred once. An example of a miscellaneous error for the graph in Figure 9c is “Men do more exercise than women and so their wellbeing is higher".

General comprehension

Viewers’ understanding of relationships depicted in graphs was poor. Perhaps the most striking finding was that only 10% of line graph readers (19% of bar graph readers) interpreted all six trials correctly. This was despite task requirements being minimal – participants were asked to spontaneously interpret the graphs to the best of their understanding and there were no time constraints. Therefore, no demanding questions were set which they needed to answer accurately or quickly. There was no association between graph format and whether all six trials were interpreted correctly (Fisher's Exact Test, $p = .66$).

Another interesting finding concerned lack of consistency in participants’ ability to interpret the different graphs presented in the experiment. Students are expected to be able to interpret material presented in graphs independent of the content or the relationship depicted, i.e., in order to be considered graphically literate,

readers need to be aware of the rules underlying graphical displays and once these rules have been learnt they should be able to apply them to these types of diagrams independent of contextual information within the graph (Friel, Curcio and Bright 2001, Shah and Miyake, 2005).

Therefore, participants' ability to interpret information depicted in graphs should remain consistent if they are accurate initially (i.e. if they interpret the first two graphs correctly the remaining graphs should also be interpreted correctly). Alternatively, if they initially provide incorrect interpretations but a learning affect occurs during the experiment and they start to interpret graphs correctly, we would expect this pattern to continue. So for example, if a participant interprets trials 3 and 4 correctly, we would expect they would continue to provide accurate interpretations. Crucially, once they start to provide accurate interpretations, we would expect they would continue to do so, rather than an erratic pattern whereby they alternate between providing a correct and incorrect interpretation.

To determine whether participants were able to provide consistent accurate responses once they had interpreted a trial correctly, I removed transcripts where participants provided correct interpretations for all six trials or no correct interpretations throughout. This left 16 verbal protocol transcripts for analysis in the line graph condition. The findings demonstrate only 31% of the sample demonstrated consistency in their interpretations. The remaining 69% were inconsistent in their interpretations of the graphs, they would interpret some trials correctly and others incorrectly in an alternate pattern. Therefore, most students were unable to reason with graphs independent of the relationship depicted or contextual information in the graph such as variable names, which vary between graphs.

This assumption was further supported by analysis of verbal protocols, for example, participants would frequently say "what does 'fractures' mean?" or "I can understand it if I focus on the stress and temperature, you can see hot temperature causes stress to increase. But I don't understand how fractures fit in." This would lead to one of the content specific errors outlined earlier, where participants focus on the two independent variables and to miss the dependent variable because they are unsure how to incorporate it into their interpretation of the graph. Interestingly, bar graph users demonstrated an identical pattern; 16 transcripts were included for analysis and of these 31% were consistent in their interpretations compared to 69% who were not.

Results – analysis of users who were not classified pre-elementary

Although the primary focus of the research concerns pre-elementary performance, not all participants were classified as pre-elementary. Therefore, another question of interest is whether those participants classified as elementary or better differ in graph reading ability between the line and bar graph conditions. If these two graph formats result in a significant difference between pre-elementary performance for line and bar graph users when readers are novices, it is possible there is a difference in other categories such as elementary and intermediate for these graph formats.

To answer this question each transcript in which participants were not classified as pre-elementary was analyzed. Similar to the previous analysis each trial was coded, but instead of the simpler coding scheme used previously where each trial was coded as either a correct interpretation or an error, each trial was analyzed based on type of information extracted from the graphs. The classification cited by Friel, Curcio and Bright (2001) was followed to categorize participants' interpretations. They reviewed criteria previous authors (Bertin, 1983; Curcio, 1987; McKnight, 1990; Carswell, 1992; Wainer, 1992) have employed to describe the type of questions graphs are used to answer. Based on these analyses; three levels of graph comprehension have been identified; elementary, intermediate and advanced.

The lowest "elementary" level of data extraction involves location of information and typically readers focus on extracting data from a graph. McKnight (1990, cited by Friel, Curcio and Bright, 2001) provides an example where the reader interprets a relationship when the answer requires them to rephrase facts, e.g., "what is the projected food production in 1985 for the developed countries?" Other examples involve readers point reading e.g., "30 cars were sold in July".

An intermediate level interpretation involves finding relationships and integrating information depicted in the graph. Wainer (1992, cited by Friel, Curcio and Bright, 2001) provides an example involving identification of trends seen in parts of the data, e.g., "Between 1970 and 1985, how has the use of petroleum changed?"

The highest level of graph interpretation is characterized by users drawing inferences from the data and considering the relationships implied by the data. Readers are required to go beyond interpreting the data to

generate hypotheses and evaluate the graph based on their quantitative knowledge. For example, extending the representation to answer a question such as "If students opened one more box of raisins, how many raisins might they expect to find?" (Curcio, 1987, cited by Friel, Curcio and Bright, 2001).

Although this classification was based on questions posed to students rather than analysis of spontaneous interpretations and tasks focussed on simple two-variable graphs, the classification was modified to use for the purposes of this research. For example, students were classified pre-elementary because they were not demonstrating the necessary interpretive skills to be classified as elementary in graph reading ability (Peebles and Ali, 2009).

Based on the classification outlined by Friel, Curcio and Bright (2001) a response was coded "elementary" if participants were reading the data. For example, in the case of the graph depicted in figure 9b an elementary interpretation may be "high protein beef results in a weight of 80 but low protein beef results in a weight gain of 90".

These types of statements were classified as elementary because participants were simply locating specific information in the graph. In order to be classified intermediate, participants need to read between the data and make inferences concerning the relationships the graph is depicting. Rather than locating specific information readers need to compare patterns to determine the general trend the graph is showing. So, for example, for the same graph an intermediate statement would involve a description of the effects the graph is depicting such as:

"Beef results in a higher weight gain than cereal irrespective of protein types, because for both high and low protein types beef is considerably higher in weight gain than cereal. However, the effect of protein type differs depending on the protein source; in beef the low protein type results in a higher weight gain than the high protein type whereas this is reversed for cereal; the high protein types result in a higher weight gain than the low protein type. "

These types of statements go beyond simply locating information and involve comparing visual chunks (for e.g., different lines) to determine whether there are any differences in the effect each independent variable

has on the dependent variable or whether the independent variables interact to influence the dependent variable.

Advanced interpretations of the graphs require application of statistical knowledge. It is not enough to simply "read the data" or even "read between the data" in these types of graphs as they are primarily designed for the purpose of data analysis. These graphs are usually used as descriptive statistics to display visually the findings of factorial experimental results. Because of this, these graphs are used so readers can visually determine whether there is a possible main effect or interaction present in the data (Pinker, 1990, Shah and Carpenter, 1995, Lewandowsky and Behrens, 1999, Kosslyn, 2006).

Therefore, in order to be classified as advanced in level of graph reading ability, participants were required to apply statistical knowledge to their interpretations. So, for example, for the same graph an advanced interpretation would involve perhaps the above intermediate description followed by an explicit identification of any main effects or interactions present. For example:

"Therefore, there is a large main effect of protein source as beef consistently results in higher weight gain than cereal irrespective of protein type. Also, there is perhaps a small interaction effect present as the effect of protein type on weight gain differs depending on the type of protein source".

The only difference between the intermediate and advanced categories was application of statistical knowledge. People in the intermediate level could be describing a main effect or interaction without having the knowledge of such concepts available in their schemas.

However, no participants in either the bar or line graph condition were classified as advanced – none mentioned main effects or interactions or how the pattern may suggest such effects. For example, expert graph readers become aware that non-parallel lines indicate an interaction effect so can identify such effects from looking at the pattern at the centre of the display (Pinker, 1990, Kosslyn, 2006). However, again no participants matched the pattern at the centre of the display to known effects, indicating limited knowledge of the graphic conventions utilized in three-variable interaction graphs.

Therefore, although a considerable number of participants were not classified as pre-elementary in performance, none of the sample was able to use the graphs for the type of analysis they are designed for – explicitly detecting effects and interactions present in the data from visually inspecting the graphs.

Analysis of transcripts from the line graph condition revealed no participants were classified as intermediate in graph reading ability. This is because no participant interpreted a minimum of four trials at an intermediate level. In the bar graph condition 10% of the sample was classified as intermediate. Therefore, both graph formats when considered together primarily result in elementary or pre-elementary interpretation of data.

Discussion

These results replicate those of the initial pilot study by Peebles and Ali (2009) and reveal that the effect of graph format on interpretation is more pronounced in an undergraduate psychology student population. The pattern of errors found is identical to that of the first study but the new results show a dramatic increase in the proportion of participants being identified as pre-elementary. In the initial study, 39% of line graph users were classified as pre-elementary. In the current experiment, the proportion of both graph users in this category increased by approximately 24% with 62% of line graph users and 24% of bar graph users being classified as pre-elementary.

Not only were the proportion of pre-elementary users and correctly interpreted trials different for the two graph types, the pattern of errors differed between the two, with line graph users being significantly more likely to ignore the x axis variable (chi-square = 6.23, df = 1, $p < .05$) or produce no coherent interpretation (missed trials) as bar graph users (chi-square = 4.27, df = 1, $p < .05$).

The findings of these experiments can be explained by the assumptions of Pinker's theoretical framework. Pinker (1990) argued information would be easy to extract from a particular graph if there were message flags in the schema specific to that information which allows for individual differences in graph comprehension. For example, an individual's schema may contain a message flag holding information about whether a line graph is depicting an interaction effect. Interactions can be identified at a perceptual level –

non-parallel lines indicate an interaction effect. Therefore, if an individual were presented with a graph where the pattern depicted non-parallel lines the message flag would be activated and they would be able to easily identify the existence of an interaction effect.

In terms of reader characteristics, an individual's graph schema may lack important message flags. Thus, they may not know that the points at the ends of each line in three-variable line graphs represent both a level of the x and z variable. If the reader is unable to extract basic information from the graph, such as which part of the visual array depicts labels on axes, they are unable to provide a full and accurate interpretation of the relationships the graph is depicting.

Pinker (1990) suggests if a reader does not have a specific schema available to them for the graph they are viewing they will rely on the closest matching schema available to them. Although three-variable interaction graphs are used for a specialist audience, the graphs are closely related to two variable Cartesian graphs. The additional complexity of three-variable graphs results from the addition of a third variable; requiring the reader to consider the interactive effects of two independent variables, each with two levels on a dependent variable.

Therefore, novices may well be approaching these graphs with interpretive schemas and processes (Pinker, 1990) for two-variable graphs. Therefore, it may not be surprising analysis of errors revealed remarkable consistency in misreading of information. When each individual error – ignoring the x, z and both content specific errors - are analyzed an overarching pattern emerges, participants are only extracting information for two variables and are unable to incorporate the third variable into their interpretation. Some participants explicitly verbalized using this strategy. Below is a participant's interpretation of the graph depicted in Figure 9c:

- "Wellbeing as a function of gender and exercise (reads title)
- That's the high and low (reading both levels of the z variable) but it doesn't tell you which ones are the males and females (levels of the x variable)
- Wellbeing, (y axis label) male, female, (x1, x2) high exercise, low exercise (z1, z2)

- Blue represents high exercise, (label colour association) which correlates with high well being = 90
- Red line is low exercise, (label colour association) which correlates with low well being
- Wellbeing as a function of gender and exercise (reads title again)
- Don't know where gender comes into it – male and female"…(unable to identify which part of the pattern depicts x value labels).

The results of this research demonstrate that if individuals are not explicitly taught to interpret these complex graphs and if they are unable to form basic associations linking the visual array to labels they are unable to provide even an elementary level interpretation of the relationship depicted in the graph. However, although reader characteristics can explain poor conceptual understanding of graphs they cannot explain differences in conceptual understanding of informationally equivalent bar and line graphs. If readers lack the message flags specifying necessary information to form a basic understanding of relationships depicted in graphs then this lack of schematic knowledge exists for both bar and line graphs which depict identical relationships and share the same framework (Kosslyn, 2006, Ratwani and Trafton, 2008).

These differences can be explained by the same Gestalt laws of perceptual organization employed earlier to account for the different IVs each group were more likely to use as the primary focus of their interpretation (Peebles and Ali, 2009). To reiterate; the sole difference between bar and line graphs is the pattern representing the data at the centre of the display.

Data points are represented in bar graphs by a single bar for each level of each independent variable with bars grouped together according to x variable value and rooted to the x axis. According to the Gestalt principle of proximity (Wertheimer, 1938) each cluster of bars forms a separate visual chunk anchored to the x axis. This ensures that when participants attend to these chunks, they are able to identify the nearby x value label quickly and easily and more readily associate the bars with the variable plotted on the x axis.

The bars are also coloured however, with a legend containing patches of the same colours next to the level labels of the other IV. According to the Gestalt principle of similarity, this shared colour allows users to also

associate each bar with its associated level rapidly and easily. The two principles combined ensure that users are no more likely to ignore one IV over another (both IVs were ignored in roughly 7% of trials).

In the case of line graphs however, data points are represented by coloured shapes (squares and circles) connected by similarly coloured lines. According to the Gestalt principle of connectedness (Palmer & Rock, 1994), each line with its two end points forms an individual visual chunk. As in the case of the bar graphs, line graph users are able to associate each line with a level of the legend variable by shared colour and the Gestalt principle of similarity.

Unlike the bar graphs however, there is no equivalent perceptual grouping process available in the line graphs to facilitate the association between the points at the ends of the lines and the variable values on the x-axis. Although points and labels may be associated by vertical alignment, it is clear that this is not sufficient to counterbalance the colour matching process, most likely because perceiving the line as the primary representational feature impairs users' ability to differentiate the points from the line.

This imbalance in the visual dynamics of line graphs results in a reduced ability of users to determine which part of the pattern depicts the variables on the x axis and in twice the number of x variables being ignored than legend variables (16.7% and 8.7% respectively). For example, for the line graph in 9d participants would often say “There is more weight gain with high protein type than with low protein type" and be unable to elaborate further or would sometimes make statements such as “There are two lines for high and low protein type but where's the information for protein source?".

The effect of the lines is more pronounced in the undergraduate population, I assume, because they have not yet acquired the interpretive knowledge that associates each point at the lines' ends with a value of both the x and legend variables. Interaction graphs are relatively uncommon and specialized compared to two-variable line graphs and in my experience many undergraduate students are encountering them for the first time in our classes.

However, despite this when students progress into further and higher education they are expected to be able to interpret graphs at a minimum elementary level but ideally at intermediate level and become advanced users after completion of education (Friel, Curcio and Bright, 2001, Mooney, 2002, Gal, 2002). These

findings demonstrate that a considerable percentage of students are below elementary level in comprehension and very few are above elementary.

Having identified the problem with line graphs, the inevitable question arises whether - and if so, how - this effect may be reduced or perhaps eliminated entirely. Three alternatives come to mind. The first is to eschew line graphs altogether and use bar graphs exclusively. Although bar graphs are currently a common choice, it has not been established that they are superior to line graphs for every task - the identification of interactions and main effects for example. Furthermore, it is by no means the case that the bar graphs cannot be misinterpreted in the same way as line graphs; 24% of bar graph users in Experiment 1 were also classified as pre-elementary.

A second way to remedy the situation is to provide explicit instruction on their interpretation and use, identifying the key representational features and contrasting them with two-variable line graphs. This avoids the more error-prone (although I suspect quite common) situation in which students must work out the rules of interpretation for different graph types through reading the literature and analyzing their own data. Although explicit teaching may be appropriate and feasible in some educational contexts, it is not always possible for all target audiences however and it is quite possible that the effect of this knowledge may diminish over time - particularly with infrequent exposure.

The most effective and widely beneficial solution therefore, may be to modify the graphical representation itself to reduce the visual imbalance and strengthen the link between the data points and all four variable values. One modification that seems - at least intuitively - plausible is to combine the features of both bar and line graphs.

More specifically, if a graphical feature like a bar were introduced to the line graph that would reinforce the connection between the line points and the x variable values (without causing additional problems or confusion through increased visual complexity), then we might predict that novice users would be less likely to ignore the x variable in their interpretations. This problem has previously been addressed by graph designers by the use of “drop lines" or “tethers" to anchor data points to reference points, lines or planes and

Harris (1999) provides a wide range of diagrams (including line graphs) with one or more such lines. In the second experiment I design a new graph with just such a modification and test this hypothesis.

Chapter 5 Modifying graphical representations

Experiment 2

The purpose of the next set of experiments is to modify the graphical display so that the visual features redress the imbalance present in the standard display. The aim is to allow the successful association of pattern to referents for novice readers whose knowledge of graphical conventions may be limited. If the modifications are successful I predict the majority of graph readers will be able to provide elementary level interpretations as well as a smaller proportion providing intermediate interpretations.

The twelve line graphs used in Experiment 1 were modified to form a set of `combined' graphs (examples of which are shown in Figure 11). In order to incorporate the bar graph feature effectively I first displaced the lines slightly (by the same distance) to the left and right so that the four line ends were placed at the same locations as the centres of the bar tops.

A dashed line from each point was then projected (of the same colour as the point) to the x axis. Dashed lines were used to reduce the perception that the resulting representation consisted of a single object consisting of two points and three lines. Compared to unbroken lines, I found that dashed lines serve to anchor the line points to the axis while maintaining the plot line as a distinct representational object. In addition, using broken lines clearly distinguishes them from the plot lines when they intersect.

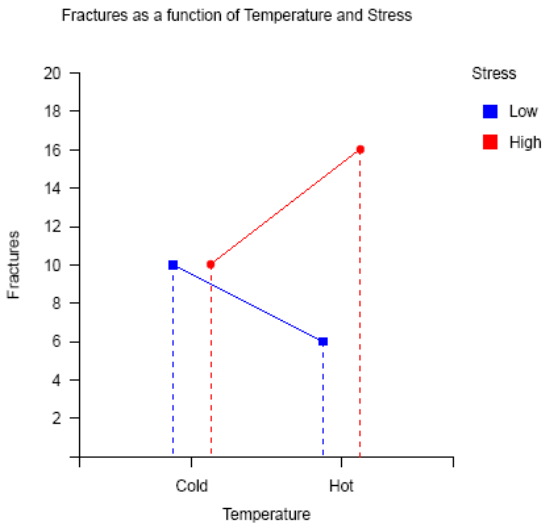
Method

Participants

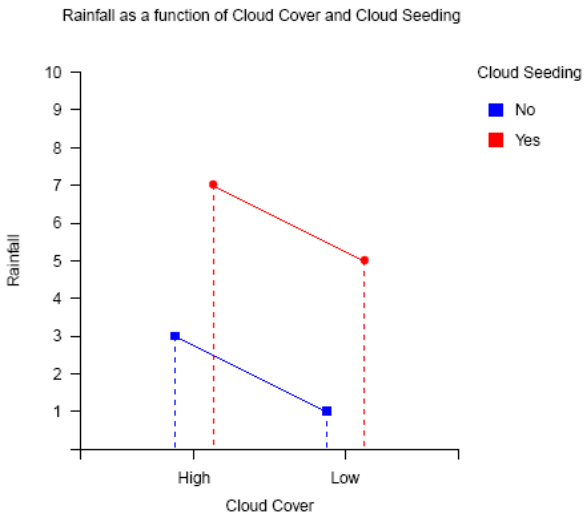
Nineteen undergraduate Psychology students (16 female, 3 male) from the University of Huddersfield volunteered to take part in the experiment, for which they were paid £5 in supermarket vouchers. The age of participants ranged from 18.3 to 44.4 years with a mean of 22.8 years (SD = 6.9). 8 participants were in their first and 11 were in their second year of study. All were alternately allocated to the experiment conditions.

Materials, Design and Procedure

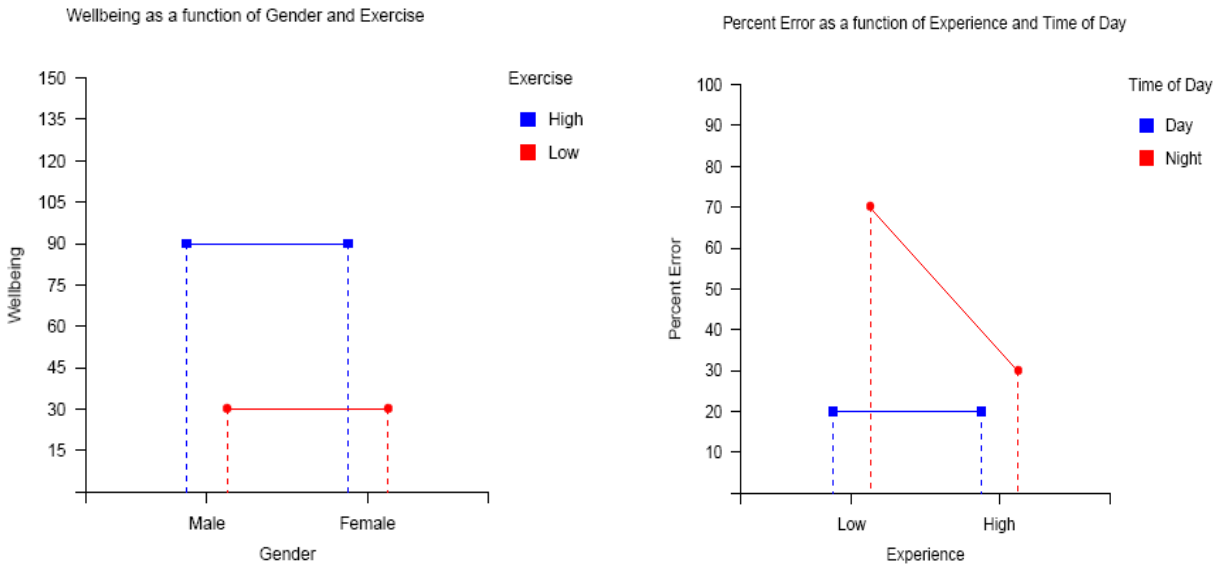
Twelve combined graphs (six normal, six reversed) were created using the same six data sets as were used in Experiment 1. The experiment was carried out using the same equipment and the same procedure as Experiment 1, the only difference being that there was only one graph condition in this experiment.



(A)



(B)



(C)

(D)

Figure 11: Four combined graphs used in Experiment 2

Results

General comprehension

There was a slight improvement in the number of participants who interpreted all six trials correctly (21% in this condition compared to 10% in the line graph condition in experiment 1). A Fisher's exact test found this association was not significant ($p = .40$). The consistency measure described in Experiment 1 revealed a similar pattern to the bar and line graph condition in Experiment 1. Removing transcripts in which participants responses were all correct or all wrong left 11 transcripts for analysis. Only 27% of the sample demonstrated consistency in their interpretation. The remaining 73% could not consistently provide an accurate interpretation after interpreting a previous trial correctly.

The data were analysed using the same method as for Experiment 1 with the authors finding a level of agreement in their coding of participants' verbal protocols of 93% ($k = .85$, $p < .001$). The proportions of erroneous and missed trials are shown in Table 3 along with those of the line graph condition from Experiment 1 for comparison. The modification produced a 25% reduction in pre-elementary performance

compared to the previous line graph condition, with only 37% of Experiment 2 participants in this category (see figure 12). Statistical analysis revealed however that this association was not significant (chi-square = 1.65; df = 1; p = .20).

A comparison of the number of correct trials in the two conditions revealed that although the combined graphs resulted in more correctly interpreted trials than the normal line graphs (mean ranks: line = 18.19, combined = 23.05) this difference was also not significant (U = 151, p = .20). Similar to the results of the earlier conditions no participants met the criteria necessary to be classified as intermediate in graph reading.

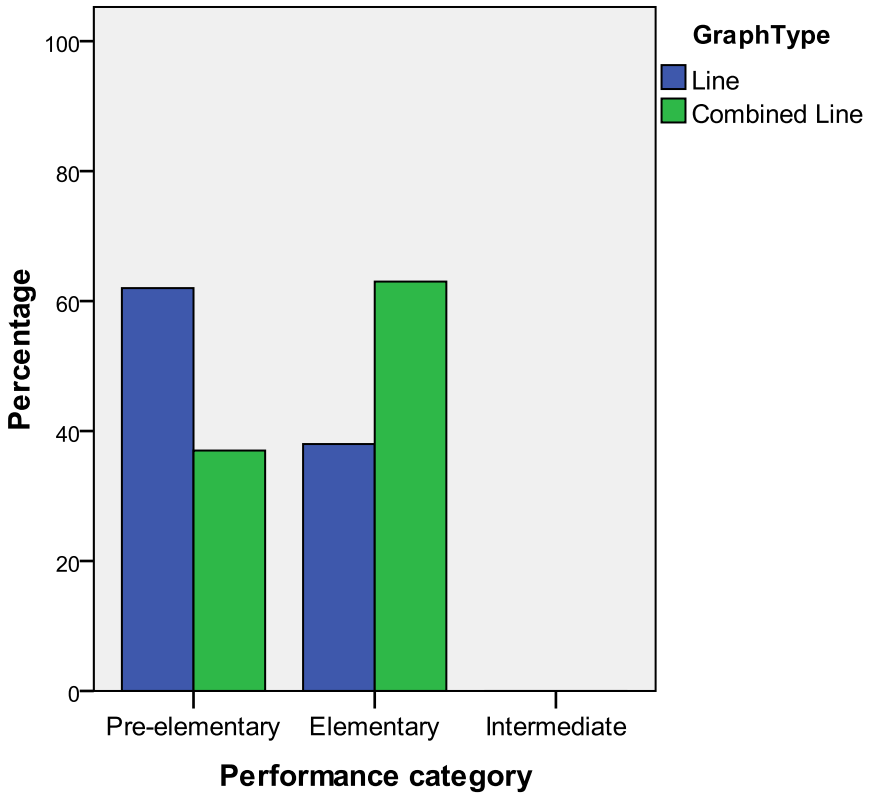


Figure 12: Percentage of line graph users in the three performance categories, combined line alongside the line graph condition in Experiment 1 for comparison.

Discussion

Although the combined graphs resulted in a reduction in the number of errors participants made, a high proportion of the sample were still pre-elementary. Consistent with the results of Experiment 1, the most common error participants made when interpreting combined graphs was to describe the effect of the legend variable but ignore the x axis variable. It seems therefore that any visual anchoring or guidance to the x axis provided by the drop lines was not sufficient to offset the salience of the coloured lines from which they project.

This may be due to the fact that colour is preattentively processed (Treisman, 1985) which draws attention early on in the interaction. Combined with the Gestalt principle of similarity, this enables a rapid and relatively effortless matching of coloured lines to legend values compared to identifying the labels at the end of the drop lines (which were displayed in the same colour as the line from which they projected to facilitate discrimination).

Analysis of the verbal protocols also revealed that participants were often surprised by the new design and unsure (at least initially) as to how to interpret the drop lines, with several commenting that they found the visual pattern confusing. Some participants asked what the dashed lines were for, or described the emergent pattern resulting from the addition of the drop down lines. For example, one participant said “not sure what this one means because of the way it is set out – it has two rectangles overlapping” (Figure 11c). Similar to Experiment 1, participants stated that they could not find the information for the x axis variable.

It is true that the addition of the drop lines, which intersect the solid plot lines, increases the visual complexity of the representation. The displacement of the plot lines slightly to the left and right of the x axis tick marks also has the effect of placing the dashed lines to either side of the x axis level labels. Unlike the bars in the bar graph, the two drop lines that project to an x axis value do not spread over the value label and do not touch. It is possible therefore that they do not combine to form an individual visual chunk with a strong link to the label in the same way the bars do.

Furthermore, it is also possible that having four lines attached to the x axis may strengthen the perception that the x axis variable is continuous. Analysis of verbal protocols indicated this was in fact the case for

some people. For example, one participant interpreting Figure 11c asked, “Does that mean males are becoming more female? I'm not sure what else it could mean" (a statement closely resembling a misinterpretation found by Zacks & Tversky (1999).

Other participants focused on the distance between the dashed lines and the label. For example, one interpretation of Figure 11d was “During the day, error was 20% and it ranges from just under low experience to just under high experience". It seems, therefore, that displacing the drop lines can not only reduce the successful association between the perceptual feature and the x axis label, but also encourage participants to attach unnecessary significance to their location.

Table 3. Percentage of erroneous and missed trials for the line graphs in Experiment 1 & 2

	Graph type	
Error	Line	Combined
Ignoring the x variable	17.46	18.42
Ignoring the z variable	8.73	6.14
Content-specific errors	8.73	6.14
Miscellaneous errors	3.97	3.51
Missed trials	9.52	7.02

Perhaps the strongest conclusion to be drawn from this experiment therefore is that although it provides some support for modifying design features to improve the base level of comprehension, the selection of which additional graphical object to introduce in a display is not trivial because factors such as visual clutter,

the strength of the visual effect introduced, and the degree of unusualness and corresponding user unfamiliarity may obviate the desired effect.

What is needed therefore is a modified graphical representation where the perceptual features relating the pattern to both independent variables are more evenly balanced. Additional constraints on any design are that it should not look too unusual or unfamiliar to users, should not over-complicate the diagram visually, and ideally should allow the same process by which readers effortlessly relate the pattern to the legend variable be employed in relating the pattern to the x axis variable.

The proposed solution to this problem is a novel design that, rather than using features that associate two locations by explicitly drawing a line between them, uses the same colour feature used for the legend variable to associate the plot points to the x axis. Examples of this new “colour match" design are shown in Figure 13.

In the new graphs, a colour patch similar to those in the legend is placed above each of the x variable values and the corresponding points at the ends of the plot lines are similarly coloured, so that, using the same colour matching process, users can more easily associate the data points with the value labels while still being able to associate them with the legend values via the coloured lines. With a more balanced representation, I predict that users will be more able to associate the data points with the values of both IVs, thereby reducing the level of pre-elementary performance to that of the bar graph condition of Experiment 1. The purpose of Experiment 3 is to test this hypothesis.

Experiment 3

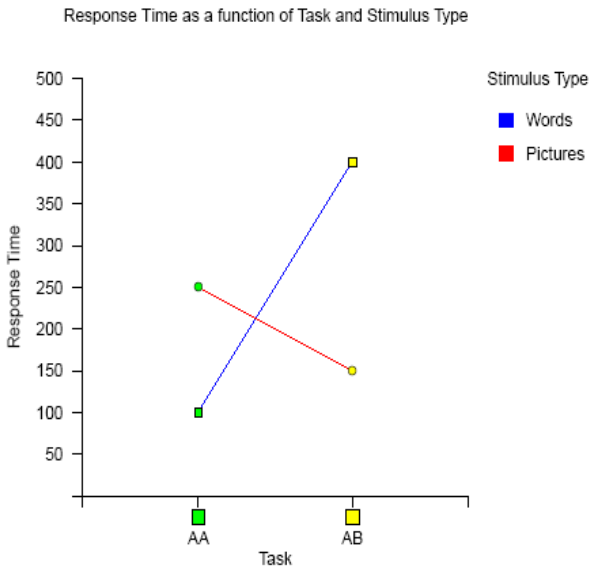
Method

Participants

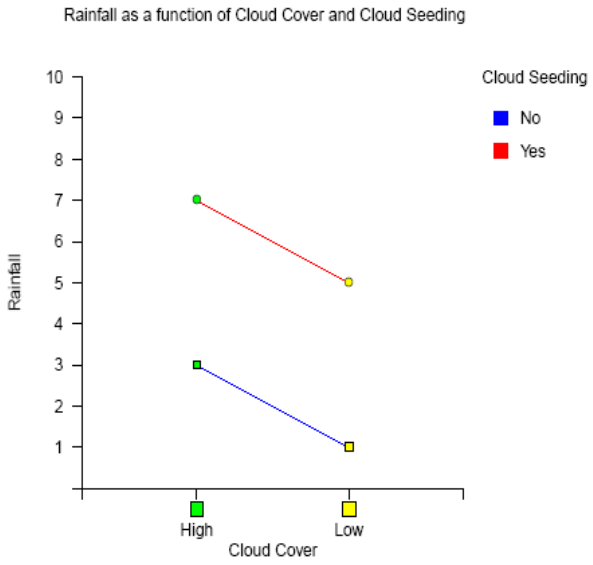
Twenty undergraduate Psychology students (15 female, 5 male) from the University of Huddersfield were paid £5 in grocery store vouchers to take part in the experiment. The age of participants ranged from 18.6 to 31.8 years with a mean of 21.8 years (SD = 3.4). Twelve participants were in their first year of study while eight were in their second year.

Materials, Design and Procedure

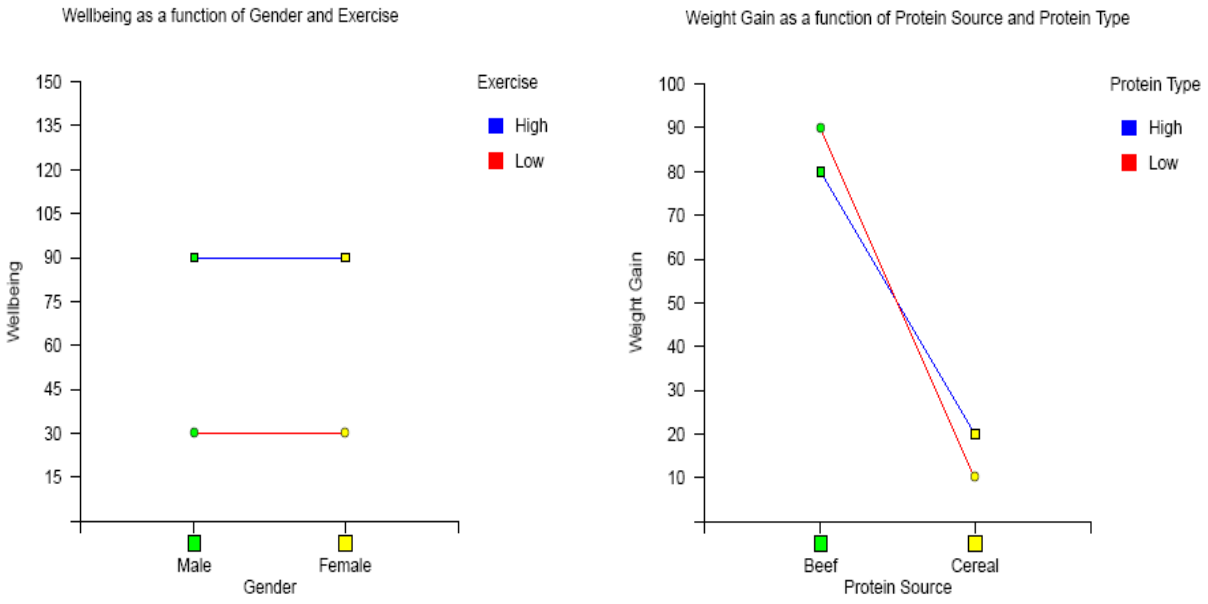
The experiment had the same design as Experiment 2, consisting of one between-subject condition: the allocation of independent variables to the x axis and legend, with 10 participants alternately allocated to each. The stimuli used in this experiment were the twelve line graphs from Experiment 1 modified to include the additional colours to the line points and the colour patches to the x axis values. Four of the stimuli are shown in Figure 13. The procedure was identical to that of Experiment 2.



(A)



(B)



(C) (D)

Figure 13: Four colour match graphs used in Experiment 3

Results

The colour match graph resulted in a number of improvements across different measures. Firstly, viewers' understanding of relationships depicted in graphs improved dramatically. Perhaps the most striking finding was that 45% of participants interpreted all six trials correctly. This compares favourably to 10% of line graph readers (19% of bar graph readers) in Experiment 1. The difference in the number of participants interpreting all six trials correctly for the colour match graph and line graph in Experiment 1 was significant (chi-square = 6.57, df = 1, $p < .05$).

Although only a small minority of the sample was categorized as pre-elementary in this condition (15%) only four participants were classified as intermediate according to the criterion outlined earlier. Therefore, 20% of the sample was classified as intermediate and the remaining 65% were classified as elementary in this condition. It would appear that the design modification was enough to aid pattern and label associations but did not then encourage the majority of students to attempt the next stage of interpreting data – identifying effects depicted in graphs.

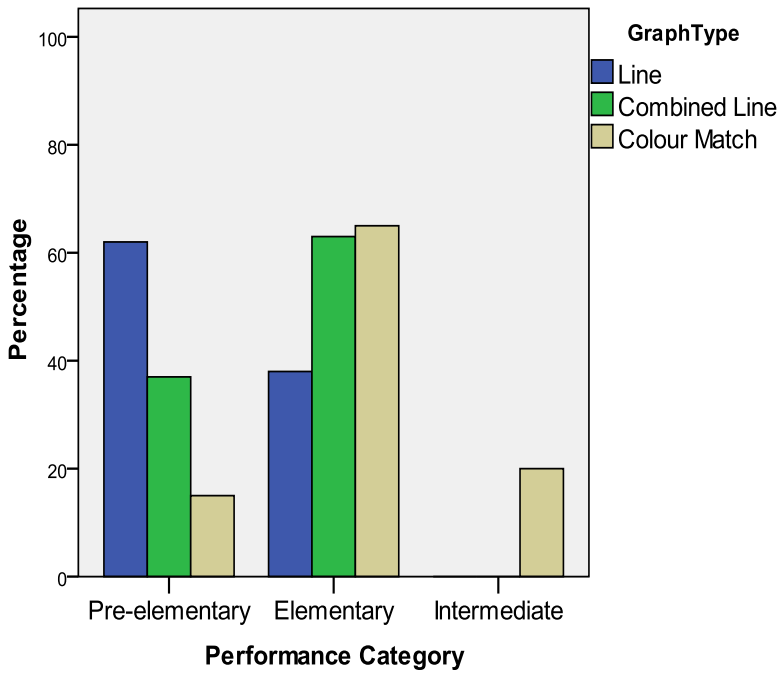


Figure 14: Percentage of line graph users in the three performance categories, colour match graph alongside the line graph condition from Experiment 1 and 2 for comparison

This is perhaps because the colour match design encourages students to point read rather than integrating information and describing direction of relationships (which allows identification of effects present in data). Analysis of verbal protocols revealed 25% of the sample predominantly focussed on point reading whilst interpreting the colour match graphs compared to 9.5% of the sample in the original line graph condition.

This could perhaps be a result of the modification implemented. The graph was modified to separate what could be perceived as a single line into individual chunks, so one line is perceived as three chunks (the line is coloured red or blue and the end points are coloured green and yellow). This modification was implemented to allow novice users to read the graphs by matching the line to the legend value and the points to each x value label through a simple process of colour matching (see Figure 13).

This resulted in four salient data points, allowing readers to easily focus on those points and provide an interpretation consisting of reading the y value for each combination of the two independent variables. For example, for the graph depicted in Figure 13a some participants would say: “Words for task Aa response time is 100 and for Ab it’s 400. Pictures for task Aa response time is 250 and for Ab it’s 150.”

Point reading differs from the more common interpretation spontaneously provided by line graph users where the trend of each visual chunk (the whole line) is described (Pinker, 1990, Shah and Carpenter, 1995, Zacks and Tversky, 1999). So for the same graph an interpretation would usually be "Response time for task Aa when the stimulus type are words is lower than the response time for task Ab. When the stimulus type is pictures Response time for task Aa is higher than the response time for task Ab" (which was a typical response provided by participants in the original line graph condition if the trial was not an erroneous interpretation).

Therefore, although the colour match design resulted in a substantial decrease in the number of readers categorized as pre-elementary, it would appear that an improvement in the ability to form basic associations between the pattern and labels does not necessarily result in readers providing a more advanced interpretation of the relationships depicted in the graphs.

Table 4. Percentage of erroneous and missed trials for the Line graphs (Experiment 1) and Colour Match graphs (Experiment 3).

	Graph type	
Error	Line	Colour match
Ignoring the x variable	17.46	6.67
Ignoring the z variable	8.73	4.17
Content-specific errors	8.73	5.00
Miscellaneous errors	3.97	0.83
Missed trials	9.52	5.00

The analysis used in the previous experiments was again employed to categorize the errors participants made with a level of agreement of 96.7% found between the two codings ($k = 0.84$ $p < .001$). The proportions of erroneous and missed trials are shown in Table 4 along with those of the line graph condition from Experiment 1 for comparison.

The modification produced a statistically significant reduction of 42% in pre-elementary performance compared to the line graphs used in Experiment 1, with only 20% of colour match graph users being classified in this category (chi-square = 7.41, df = 1, $p < .01$). Although this figure also represents a performance improvement of 17% compared to the combined graphs of Experiment 2 and 4% compared to the bar graphs of Experiment 1 (see figure 14) these were not statistically significant (combined: chi square = 1.37, df = 1, p = .24; bar: Fisher's Exact Test, p = .53 (one tailed)).

A comparison of the number of correct trials between the conditions also revealed that the colour match graphs resulted in a significant increase ($H = 9.33$, df = 2, p = .03) in the number of correctly interpreted trials (mean rank = 51.98) compared to the normal line graphs (mean rank = 25.74), combined graphs (mean rank = 31.2), and bar graphs (mean rank = 36.07).

Three post-hoc Mann Whitney U tests (with alpha levels Bonferroni adjusted to .017) revealed the significant difference to be between the colour match and line graph condition ($p < .01$), but not between the colour match and bar graphs (p = .18) nor between the colour match and combined graphs (p = .07).

As with the previous experiments, there was no significant association between performance and year of study (Fisher's Exact Test, p = .15 (one tailed)), nor by whether they saw `normal' or `reversed' graphs (Fisher's Exact Test, p = .29 (one tailed)).

Discussion

In producing such a significant reduction in pre-elementary performance, the colour match design supports the suggestion that standard line graphs create an unbalanced visual representation which over-emphasizes the legend variable values to the detriment of the x axis ones. The results of Experiment 3 also support the hypothesis that additional colour patches are sufficiently salient to balance the representation by drawing users' attention to the x axis values without looking too unusual or unfamiliar to users or making the diagram too visually complex.

Figure 15 displays the error rates for all four graph types together. It shows that the colour match graphs produce the lowest number of errors of all the graphs. Crucially, the pattern revealed in the previous experiments - that readers are twice as likely to ignore the x axis variable as they are the legend variable - was not found. In this condition the frequencies of these two errors were much closer.

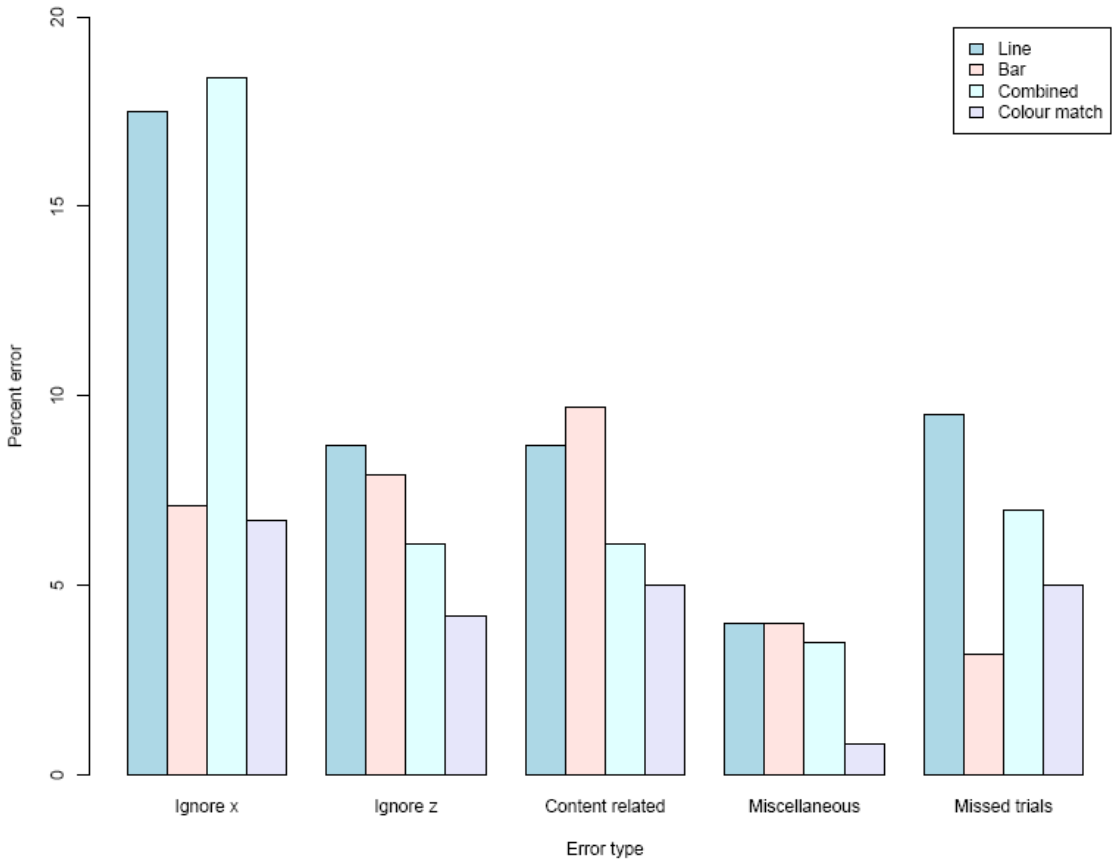


Figure 15: Percentage of errors by error and graph type, Experiments 1-3

This pattern can be explained by identifying how many- and which – Gestalt organization principles are having an effect. In the original line graph condition, the principle of similarity allowed participants to relate plot lines to legend values by colour, but there was no equivalent grouping principle facilitating the association of plot features to the x axis values.

In fact, this association is actually hindered by the operation of a second Gestalt principle; connectedness (Palmer & Rock, 1994) which encourages the perception of plot lines as single objects rather than as connections between data points (Pinker, 1990, Zacks and Tversky, 1999). This combination of Gestalt principles strongly directs novice users to relate the plot pattern to only the legend and y axis variables, resulting in the catalogue of errors found in the previous experiments

In the colour match graphs, differentiating the plot lines from their data points by colour prevented participants from perceiving the line as a single object and made the individual data points more visually salient. Placing the colour patches above the x axis values then balances the visual dynamics of the graph by bringing the Gestalt principle of similarity into effect for the x axis variable as it does for the legend variable - readers can match the line colours to the legend values and the data point colours to the x axis values.

This analysis is supported by the verbal protocols recorded. In the previous experiments participants would often match plot lines to legend values, (e.g., for Figure 13c “Blue is high exercise, red is low exercise") but then fail to incorporate the x variable values into their interpretation. Users of the colour match graphs however, were far more likely to continue their interpretation of Figure 13c, e.g., “Blue is high exercise, red is low exercise. Green is male and yellow is female". By allowing novice readers lacking the interpretive knowledge for these graph types to associate all referents to the plot pattern using the same visual features and Gestalt principles, the colour match design balances the features of line graphs and brings user performance on a par with that of the bar graph users in Experiment 1.

General discussion

Gestalt principles are an important factor in the visual processing of graphical representations. Pinker (1990) for example, argues that Gestalt Laws of Perceptual Organization (Wertheimer, 1938, Palmer and Rock, 1994) are one of the four key principles that determine the nature of the mental representations that users generate when reading a graph. According to Pinker, the Gestalt laws of proximity, similarity, connectedness, good continuation and common fate all determine how individual graphical features are grouped together to form coherent wholes and so relate patterns to variables and their values together.

Pinker cites research showing that Gestalt principles can be combined to facilitate comprehension. Parkin (1983, cited in Pinker, 1990) manipulated the number of Gestalt principles associating labels to lines in a line graph in order to ascertain how this affected comprehension. He compared the speed of readers' comprehension times to graphs with labels utilizing no Gestalt principles (placed in a legend or a caption) to labels with one Gestalt principle (proximity, good continuation or similarity) and two Gestalt principles (proximity and good continuation). Consistent with predictions, it was found that providing principles did not lead to a competing organization of labels with labels increasing the number of Gestalt principles associating labels to lines led to a reduction in response time.

Shah, Mayer, and Hegarty (1999) have also demonstrated how the appropriate use of Gestalt principles can improve the interpretation of statistical graphs. They conducted an experiment identifying graphs from social science textbooks which high school students failed to interpret appropriately (the students did not describe the overall trends the graphs depicted but simply focussed upon specific values). The authors argued that this was due to inappropriate grouping of perceptual information in the graphs rather than the graph format used and, using Gestalt principles, they regrouped the relevant information, either by connecting data points in a line graph (the principle of connectedness) or by placing them together in bar graphs (the principle of proximity). The modified graphs significantly increased the ability of students to identify the global trends in their interpretations, demonstrating that, when used appropriately, Gestalt principles can improve conceptual understanding of statistical graphs.

Kosslyn, (1989) also regards these Gestalt principles as being vital in determining the ease with which graphical representations can be understood. Kosslyn suggested a set of “acceptability principles" for the various components of a graph which he argued must be followed in order for it to be read appropriately. For example, he advises that variable labels must be sufficiently close to the feature representing the variable (relative to other features), in order for the Gestalt principle of proximity to operate so that readers can easily associate the two.

A negative consequence of this relationship however, is that, if care is not taken in the design of a graph, Gestalt principles may group elements inappropriately, leading to failures in comprehension. For example, Kosslyn (1989) illustrates this point with a Cartesian graph in which the y axis label is placed too close to the origin. Kosslyn (1989) argues that this violates his acceptability principle of “organization of framework and labels" because the label's proximity to both x and y axes makes association ambiguous. This can be remedied by explicitly positioning the label closer to the vertical scale.

While no doubt true that the relationship between Gestalt principles and comprehension can have negative consequences if not appropriately applied, as Lewandowsky and Behrens (1999) have argued, producing guidelines for avoiding these limitations is problematic due to there being no accepted principles for predicting what constitutes inappropriate grouping in statistical graphs.

For example, although I have highlighted negative consequences of the Gestalt principle of connectedness operating in line graphs, it is this very same principle that allows experienced readers to integrate data and identify trends (Schutz, 1961) or rapidly interpret frequently encountered patterns. A prime example of the latter in the 2 x 2 interaction graphs used in these experiments is the cross pattern (Kosslyn, 2006), an example of which is shown in Figure 9a. Experienced graph readers can often swiftly identify this pattern as representing a “crossover interaction" between the two IVs and explain that it reveals that they are not independent but that pairwise combinations of their levels produce reversals in relative DV values.

Such considerations have led researchers to stress the importance of taking into account the specific requirements of the intended task and how well they are supported by the representational properties of different graphical features when deciding which graph format to use (e.g., Peebles & Cheng, 2003; Peebles,

2008). Task and graphical representation are only two dimensions of the cognition-artifact-task triad (Gray & Altmann, 2001) however, and it is also vital to understand the characteristics of the various intended users of the graph.

One of these characteristics is domain knowledge and a number of studies have shown that users' interpretations of graphical representations can be affected - for good and ill – when they have some knowledge of the variables and how they relate to each other. For example, it has been demonstrated that people are more likely to extract general trends in line graphs and misestimate correlation strength in scatter plots when the variables are known compared to unfamiliar ones (Shah and Hoeffner, 2002; Freedman & Smith, 1996). Shah (1995) has also shown that domain knowledge can cause novice graph users to interpret relationships incorrectly if the positioning of variables does not follow convention (i.e., if the axes representing the DV and IV are reversed).

A small subset of errors for two graphs in these experiments are interpreted as resulting from participants' prior knowledge of the relationships between the variables - specifically the relationships between temperature, stress and fractures (graph 2) and between protein type, protein source and weight gain (graph 6). In both cases these content-related errors were relatively rare and were found in both graph conditions. However, in comparison to the number of non-content related errors this study has revealed, the effect of content on interpretation can be seen to be relatively minor. These studies show that, for novice users of 2 x 2 interaction graphs, the effect of graphical representation far outweighs that of content.

As part of their training students of the natural and social sciences are expected to develop sophisticated graphical literacy skills as much of their work will involve the production and interpretation of graphical displays of data. Interaction graphs form a significant proportion of this experience and it is vital therefore that the processes involved in their use are understood so that skills may be taught appropriately and the best graphical formats used.

Students' difficulty with interaction graphs may, in part, be due to the coverage of them in the statistics textbooks they encounter during their studies. In discussing graphical representations of factorial designs, statistics textbooks aimed at undergraduate psychology students either focus entirely on, or strongly

emphasize, the interpretation of main effects and interactions (e.g., Howitt & Cramer, 1998; Aaron et al., 2006; Dancey & Reidy, 2008; Field, 2009; Langdridge & Hagger-Johnson, 2009).

While this is not surprising given that this is the primary function of such graphs, it may often be the case that students are being presented with advanced interpretive instructions while their basic conceptual understanding of the graphical representation is lacking. This research suggests that students' difficulties with these graphs could be addressed by more explicit instruction on the basic representational features of interaction graphs and the processes required to interpret them correctly.

It has been assumed that students can interpret both bar and line interaction graphs equally well and that the benefits of line graphs enjoyed by experts can readily be acquired by novices. I have demonstrated the limitations of this assumption and shown that a large proportion of undergraduate students struggle to interpret line graphs even at an elementary level. Although the use of bar and line graphs is roughly equal in the research literature, it may be the case that students have greater exposure to line graphs because of the textbooks and statistical software they use.

There are several possible responses to these findings. One is to maintain the status quo, continue to employ both bar and line graphs equally with the recommendation that the correct interpretation of line graphs be more explicitly taught. While this is indeed an option, it is limited because it places the onus of successful interpretation on external factors, thereby risking the possibility that it may not be carried out appropriately, for example due to lack of space for detailed instruction in a curriculum.

Another response is to suggest that students be encouraged to use bar graphs predominantly and recommend that bar graphs be more widely used in textbooks and research literature. While I regard this approach as perhaps being a more practical and viable option than the previous one, it too is limited. A consequence of adopting this approach would be that students receive less exposure to line graphs and so are less likely to acquire the pattern recognition schemas that experts use so effectively.

A third alternative is to adopt the colour match graph I have developed here which combines the benefits of both line and bar graphs. Students using this graph format would benefit from the balanced visual dynamics found in bar graphs which facilitates the matching of data points to the levels of both IVs through colour,

while maintaining the global line-based patterns found to be so useful in line graphs. This design-based solution provides the appropriate representational features to support correct associations between pattern and referents which promotes accurate interpretation and the development of pattern recognition schemas.

Chapter 6

Honours level students

This experiment will develop the findings from the previous experiments further. The previous experiments were conducted in an educational context so the research findings could be applied to student learning. However, the sample used for the first three experiments consisted of first and second year undergraduate students only. This precautionary measure was taken to ensure varying levels of exposure and teaching did not confound the experimental results and give those with advanced training an unfair advantage. However, there was no significant difference in pre-elementary categories between the first and second year students and so the results were amalgamated together.

Excluding final year students close to graduating leaves open the question of whether these students (who have had a considerable amount of exposure to these types of graphs from the educational and research literature as well as the training received in research methods modules) develop greater expertise in handling quantitative information than first and second year students who are still undergoing their training.

To address this question a further experiment was conducted with final year undergraduate students used as the sample. To ensure students had benefited fully from the training in quantitative research the experiment was conducted towards the end of the academic year. At this time students had completed most of their assessments (apart from exams) and the final teaching term was finishing. Only those students who had received training in core quantitative research methods modules at foundation and intermediate level were included as part of the sample.

Therefore, the aim of this experiment was to investigate differences in graphical literacy skills between undergraduate students early in the course and students close to completing their degree. This specific sample was used to determine whether pre – elementary performance is a function of stage of study. If a large proportion of third year students are categorized as being pre – elementary then it would be reasonable to assume that students who will soon be graduates have a poor level of graphical literacy despite a high degree of exposure to interaction graphs throughout their studies. If this is the case then students are

graduating lacking one of the skills required to be defined quantitatively literate (Friel, Curcio and Bright, 2001).

Experiment 4

This experiment was a replication of Experiment 1 except final year undergraduate students were used as the sample.

Method

Participants

Twenty-nine undergraduate Psychology students (24 female, 5 male) from the University of Huddersfield volunteered to take part in the experiment, for which they were paid £5 in supermarket vouchers. The age of participants ranged from 20.7 to 31.3 years with a mean of 21.9 years (SD = 2.2). All participants were in their third year of a three-year psychology degree.

Design

The experiment was an independent groups design with two between-subject variables, type of diagram used (bar or line graph) and the allocation of independent variables to the x axis and legend (labelled `normal' and `reversed'). Fourteen participants were allocated to the bar graph condition and 15 to the line graph condition. There were 7 participants in the normal-bar condition, 8 in the normal-line condition, 7 in the reversed-bar condition and 7 in the reversed-line condition.

Materials and Procedure

The materials and procedure were identical to Experiment 1

Results

The data were analyzed using the same method as Experiment 1, to categorize trials into correct or erroneous interpretations. Data was analyzed to determine whether the bar – line difference found in Experiment 1 was replicated with final year students. The bar-line difference emerged again: 60% of line graph users were classified pre-elementary compared to 7% of bar graph users with the third year sample. Statistical analysis

revealed that this association between line graph users and pre-elementary performance was significant (chi-square = 8.96; df = 1; $p < .01$).

A comparison of the number of correct trials between the two conditions revealed that the bar graphs resulted in more correctly interpreted trials than the line graphs (mean ranks: bar = 19.21, line = 11.07); this difference was significant ($U = 46.00$, $p < .01$).

Table5: Percentage of erroneous and missed trials, Experiment 4

	Graph type	
Error	line	Bar
Ignoring the x variable	22.22	4.76
Ignoring the z variable	11.11	3.57
Content-specific errors	12.22	5.95
Miscellaneous errors	2.22	1.19
Missed trials	3.33	1.19

The pattern that emerged in this experiment was highly consistent to the results found in Experiment 1 (see figure 16). Like the first and second year students still undergoing their training in research methods final year students who had completed their training were still poor at extracting the relationships depicted in the graphs. Although final year students are not explicitly taught research methods, they receive exposure to these types of graphs from reading the literature and conducting analyses for their own research project. Similar to Experiment 1 only 13% of line graph readers interpreted all six trials correctly compared to 36%

of bar graph readers. The difference in the number of participants interpreting all 6 trials correctly in the bar and line graph condition was not significant (Fisher's Exact Test, p = .215).

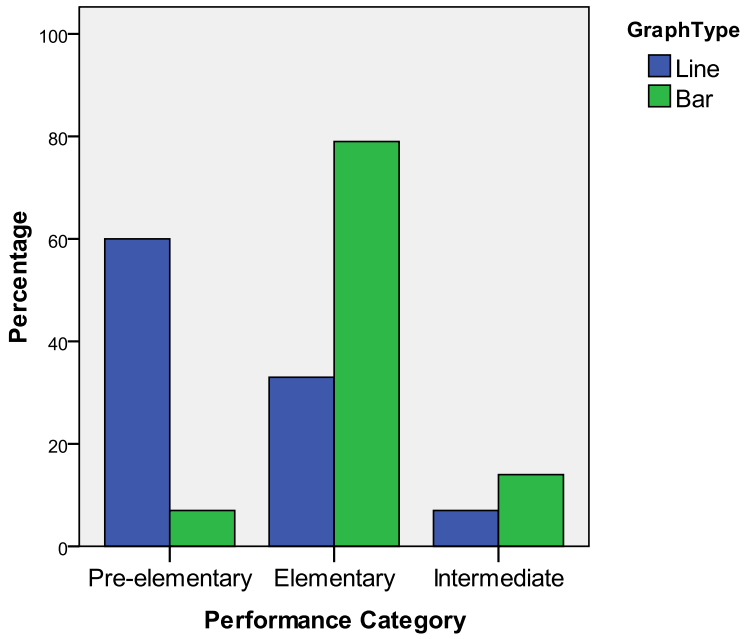


Figure 16: Percentage of bar and line graph users in the three performance categories.

Discussion

The high rate of pre-elementary performance found in this experiment demonstrates that many students are graduating without the ability to interpret a key graph format appropriately. Increased exposure to such displays from reading the research literature, textbooks and training in data analysis had no marked effect on the interpretations provided, especially in the case of line graphs. The findings of this experiment effectively replicate those of Experiment 1.

Although there has been previous research in the area of graph comprehension investigating differences between bar and line graphs, this research has typically focussed on the type of data these two informationally equivalent graph formats are appropriate to depict. In a series of studies Zacks and Tversky (1999) investigated the well known bar – line correspondence where individuals are better at making discrete

comparisons using bar graphs and trend comparisons using line graphs (Simcox, 1984, described in Pinker, 1990, Pinker, 1990, Carswell and Wickens, 1995, Kosslyn, 2006).

Similar to Simkin and Hastie's (1987) research Zacks and Tversky (1999) asked participants to provide spontaneous interpretations of bar and line graphs. They found that viewers overwhelmingly gave discrete interpretations of bar graphs but interpreted line graphs as depicting continuous trends. When participants were asked to construct graphs from statements provided the same pattern emerged, if viewers were given a statement involving a discrete comparison they tended to construct bar graphs whereas those participants who were given continuous statements tended to draw line graphs.

Zacks and Tversky (1999) explain their findings and those of previous research investigating the bar-line correspondence as emerging from cognitively natural ways of using space to convey meaning. The Gestalt principles underlying figural perception support the naturalness of bars for categorical information and lines for continuous data (Pinker, 1990). In bar graphs each label value is represented as a separate bar suggesting separate entities or categories, whereas in line graphs values are connected by a single line suggesting that all the values belong to the same entity. These assumptions are further supported by cross – cultural research. Children across cultures line up dots they perceive as representing levels of an underlying dimension but do not line up dots they do not perceive as related dimensionally (Tversky, Kugelmass & Winter, 1991).

Secondly, there has been previous research into students' understanding of graphs in educational settings but this research has typically focussed on elementary and middle school students to determine whether they can adequately read graphs (Curcio, 1987, Preece and Janvier, 1992, Phillips, 1997). There is an assumption in the literature that once students progress into further and higher education the core skills needed to interpret Cartesian graphs are available to them (Curcio, 1987, Friel, Curcio and Bright, 2001).

Furthermore, research in such settings has focussed on predictors of graph comprehension. For example, Curcio (1987) looked at whether sex, prior knowledge of the topic, form of the graph and reading and mathematics achievement affected students' scores on a graph comprehension test. He found that at grade four all measures apart from sex were predictors of graph comprehension. By grade seven however, prior knowledge of topic and graph format did not predict graph comprehension scores.

Of particular relevance to the current research, graph comprehension research in educational settings has found students find line graphs more difficult to comprehend than other graphs (Bell, Brekke, & Swan, 1987). Culbertson and Powers (1959) suggested line graphs are more difficult to comprehend because of the sparseness of information and more abstract representation. They concluded bar graphs are easier to read than line graphs. Consistent with the explanation I proposed, they suggested the reason bar graphs are easier to read is because they clearly connect the horizontal axis with each abscissa value to be read. Line graphs do not clearly pinpoint abscissa values so picking out points on a line may make comprehension difficult. However, their task partially consisted of asking participants to compare specific quantities, later identified as a task appropriate for bar but not line graph displays.

Furthermore, in their review of the literature of factors influencing graph comprehension Friel, Curcio and Bright (2001) made recommendations for when graphs should be introduced to students in the classroom. In their classification bar graphs can be introduced as early as Key Stage 2 whereas line graphs should be introduced later, between Grades 6-8. At this stage students are expected to be able to comprehend and construct line graphs because of increased sophistication in abstract reasoning ability which research has demonstrated is necessary to interpret line graphs (Dillashaw & Okey, 1980; Padilla, McKenzie, & Shaw, 1986; Berg and Phillips, 1994).

Research into differences in students' ability to read bar and line graphs tapers off once the sample consists of further and higher education students. This is because the imbalance in understanding these two commonly used graphs in earlier education appears to become balanced whilst students are still in middle school (Friel Curcio and Bright, 2001). However, these findings concern simple Cartesian graphs, usually those depicting the relationship between two variables.

The bar –line difference found in the experiments reported here and by Peebles and Ali (2009) is a robust finding replicated over numerous experiments. Students find it easier to read information depicted in bar charts than if the same information is plotted in a line graph when task requirements are controlled for. I attribute this finding to the complex graphic conventions present in three-variable graphs which are not present in two variable graphs.

For example, in contrast to two-variable Cartesian graphs which consist of an L shaped framework plotting two variables, three-variable graphs include an additional third variable usually plotted in a legend. In order to provide an elementary level interpretation of data depicted in these graphs all three variables need to be taken into consideration. Students are required to understand that the x and legend values are independent variables but can sometime interact to influence the dependent variable. In addition to this, students need to realize that the line graph display uses the endpoints of the lines to depict each level of the two independent variables.

The results of these experiments reveal that students are unaware of these graphic conventions and perhaps the reason they do not pick them up from exposure to these types of graphs is because teaching material does not incorporate explanations of how to decode these types of graphs. Educational textbooks include these interaction graphs throughout the text implicitly assuming students will know how to interpret them. Even statistic books, for example Dancy and Reidy (2004) which include explanations of how to interpret graphs focus on higher level data extraction (e.g., identifying main effects, interactions) but fail to include instruction concerning basic interpretive processes involved in graph interpretation (e.g., knowing the x and z variable are independent of each other). Presumably this is because authors of these texts assume basic level interpretive knowledge concerning graphic conventions is available to readers. This mistaken assumption is understandable however as this is the first piece of research to demonstrate that the large majority of students cannot interpret these types of graphs at an elementary level.

General discussion

Statistics and quantitative research methods, core skills taught in social science degrees, involve a heavy reliance on graphical representations as a tool to analyze data. One such graphical representation is the three-variable bar and line interaction graphs used in the experiments reported here. The relationships between variables these graphs can communicate from visual inspection of the display make them a powerful tool for analyzing data at the initial exploratory analysis stage (Kosslyn, 1994, Friel, Curcio and Bright, 2001). In particular, when data is plotted in the line graph format the pattern formed by the lines allows expert readers

to identify effects present in the data rapidly and easily by visual means (Pinker, 1990, Shah and Carpenter, 1995, Kosslyn, 2006).

Students in the social sciences undergoing training in quantitative research methods are expected to be able to analyze such data with a high degree of sophistication once their training is complete. However, the experiments reported here and in the Peebles and Ali (2009) paper clearly demonstrate participants struggle to interpret these graphs at an elementary level and provide erroneous interpretations on a number of trials.

Furthermore, for those who are accurate in their interpretations few can provide consistent intermediate interpretations. Although the design modifications in Experiments 2 and 3 reduced rate of pre-elementary performance, few participants were classified as intermediate graph readers. This was even the case in Experiment 4 where the majority of participants provided correct interpretations in the bar graph condition and pre-elementary performance dropped to 7%. Therefore, although final year students performed well in the bar graph condition by correctly interpreting the graphs few went beyond descriptively reading graphs to compare the differential effects of the independent variables, which would have been classified as an intermediate interpretation.

Perhaps more striking is the finding that across all four experiments - totaling 110 participants - who ranged from being in the first year of a three year psychology degree to final year students close to graduating none mentioned the effects graphs were designed for – simple, main and interaction effects. The large nature of the sample (when the experimental results are combined together) suggests these concepts are not available in undergraduate students' graph schemas. For example, Pinker (1990) points out a reader may lack important message flags so they may not know that sloping lines indicate an interaction effect.

This assumption is supported by verbal protocol data – for example, some participants would say in their interpretation: "the lines cross but I don't know what that means" (figure 9a). Furthermore, when we consider the overall level of interpretations provided by participants the conclusion that students do not have these advanced skills is not surprising. Over half the line graph condition in both Experiment 1 and 4 were classified as pre-elementary. Some participants explicitly stated they did not know where information for a particular variable was in the display.

If students do not have schematic knowledge allowing them to form basic pattern and label associations it is not surprising that more advanced knowledge of graphic conventions (how to determine whether there is an interaction or main effect present) is also unavailable to them. Therefore, the primary purpose of these types of displays is not benefiting undergraduate students.

Pinker (1990) suggests formal instruction (providing students also have the opportunity to construct graphs themselves) can enrich graph schemas so they contain necessary and sophisticated message flags that allow an individual to interpret a graph at an advanced level. The results of the experiments conducted so far suggest that formal instruction is necessary; although the design modifications implemented in Experiment 2 and 3 reduced pre elementary performance few readers could be classified as intermediate and none were advanced. Because the sample consists of higher education students it is imperative students are able to interpret quantitative information depicted in these types of graphs ideally at an advanced level.

Chapter 7

The different effects of thinking aloud and writing on graph comprehension

The focus of the experiments up to this point has been assessing students' conceptual understanding of graphs and whether it is possible to modify diagrams such as these to improve basic processes involved in graph comprehension such as associating data points to referents. To investigate this research question the verbal protocol method was employed.

This method was employed to uncover what features of the representation produced the errors observed in Experiment 1 and the high rate of pre-elementary performance observed in the line graph condition. Although other methods would have allowed me to record students' interpretations of these graphs (e.g., question answer tasks, drawing tasks) they do not allow researchers to trace the underlying cognitive processes leading to the errors students make whilst attempting to provide an interpretation (Crutcher, 1994, Payne, 1994). However, the think aloud method is not necessarily an accurate reflection of how students interact with educational material. The experimental conditions in the experiments conducted so far for this research require students to report their thoughts continuously whilst undergoing the task which does not necessarily accurately reflect how students interpret such information when presented with it.

As this research is applied to educational learning it seemed appropriate to investigate whether requiring participants to write their response (as opposed to thinking aloud) would result in any difference in interpretation provided. Students are often required to include these graphs in reports presenting results of factorial research designs and include a written interpretation of the results the graphs are depicting. A comparison of these two methods will hopefully help ascertain the appropriate learning strategy for students to employ whilst attempting to understand educational material. If some learning occurs whilst writing an interpretation of these graphs I can recommend students write an interpretation of graphs when they see them in textbooks, journals or other educational material. However, if verbal protocol responses are found to be superior to written responses, I can recommend students think aloud whilst interacting with such data.

There is a large body of literature investigating whether writing improves conceptual understanding of material in a number of disciplines (e.g., Britton, 1978; Flower & Hayes, 1980; Young & Sullivan, 1984; Newell, 1984; Bereiter & Scardamalia, 1987). This research comes under the umbrella of "writing to learn" and advocates assert that writing can help engender critical thinking and the formation of new relationships between ideas, leading to knowledge construction (Klein, 1999). Several processes involved in writing have been identified as possible causes for these observed improvements in conceptual understanding. For example, the self-paced nature of writing allows for reflection (Emig, 1977; Ong, 1982) while the permanence of the text allows material to be reviewed (Emig, 1977; Young & Sullivan, 1984). The process of reviewing allows the writer to judge what is written against what is intended to be communicated and to evaluate (and improve) the logical coherence of sets of sentences within the text (Galbraith, 1992).

Furthermore, the context in which writing is produced can result in improved conceptual understanding of material. For example, the absence of an immediate audience requires writers to be explicit in their interpretation and presentation of material (Olson, 1977). In a review of the evidence into the effect of writing on learning however, a number of authors have concluded the evidence to support the above assertions is lacking. Klein (1999) concluded that the evidence indicating that writing improves conceptual understanding of material is inconsistent. For example, in an influential earlier review Applebee (1984) noted the research studies conducted to answer this question lacked control groups and implementation of pre and post tests. Based on these findings Applebee (1984) concluded that this research question was lacking rigorous investigation.

Ackerman (1993) reviewed 35 studies from the writing to learn literature and concluded that they failed to present evidence that writing results in learning. This led him and a number of other authors to criticize the writing to learn model and conclude that the act of writing itself does not result in improvements in learning (Sensenbaugh, 1989, Schumacher and Nash, 1991, Rivard, 1994). A more recent review of the literature led Kline (1999) to conclude that although the evidence for the assertion that writing produces positive learning effects is stronger than the period during Applebee's (1984) review the actual findings are mixed. Research papers have found diverging results ranging from positive, negative to no effects making it difficult to conclude whether writing has any instructional value in its own right.

One example of the positive effects of writing is a study by Benton, Kiewra, Whitfill and Dennison (1993) investigating whether note taking can result in improvements in essay writing. Undergraduate psychology students watched a tape about various forms of creativity. Students were either instructed to take notes or to simply watch the tape. Participants then wrote an essay comparing different types of creativity. Benton et al (1993) found that those participants who had written notes wrote lengthier and more organized essays than those who produced no notes.

Tynjälä (1999) explains these conflicting findings as resulting from differing tasks demands. If writing involves low level learning such as accumulation of factual knowledge then writing will result in no difference to a passive method such as reading material (Penrose, 1992). However, when higher-order thinking is required writing can result in learning gains. Tynjälä concludes that generally writing is an effective learning tool when attempting to advance students' understanding and critical thinking skills but not superior to any other method when students are required to simply "tell what they know". In a similar vein, the second factor that can explain the conflicting results is how much information manipulation is required from the task. The larger the demands of manipulation of information are the stronger the learning effects should be (e.g., Applebee, 1984; Langer, 1986, Greene & Ackerman, 1995; cited by Tynjälä, 1999).

In relation to the current research question it is difficult to ascertain whether writing would improve students' understanding of the graphs they are required to interpret. The literature investigating the effects of writing on learning compares different writing activities (e.g., writing weekly reports to keeping a journal) or a writing condition to a control group (no writing) or comparing writing to other study behaviours, e.g., reading (Penrose, 1992, Ackerman, 1993, Greene, 1993, McCrindle and Christensen, 1995) rather than comparing different methodologies.

In addition to this, the verbal protocol method has been widely adopted in the writing to learn literature (Flower and Hayes, 1981, Hayes and Flower, 1981, Cumming, 1989, Greene, 1993) in an attempt to uncover cognitive processes involved in writing. Therefore, any potential benefit either method can provide may be confounded by the simultaneous use of both methods to study cognitive processes involved in writing.

In contrast to the writing literature where the argument is that writing improves comprehension of material according to Ericsson and Simon's (1993) theory of protocol generation the act of thinking aloud concurrently during a task should neither impair nor enhance performance as participants are simply verbalizing their thought processes. The think aloud method allows access to participants' short-term memory stream and verbalizations uncover cognitive processes involved in task completion but do not alter them when Type 1 or Type 2 verbalizations are employed.

Previous studies have demonstrated that undergraduate university students' ability to understand statistical data can vary significantly depending on the form of the graphical display. Specifically, this research has shown that for a considerable number of students, conceptual understanding of three variable line graphs does not meet the lowest level of graph comprehension ability identified in the literature. This finding led me to propose a fourth, lower category of comprehension ability which I termed "pre-elementary" and subsequently to propose and test a novel line graph design which I found successfully reduces the error level to that of the bar graphs (Experiment 3).

Developing an adequate model of diagrammatic reasoning requires taking into account three interacting factors: the nature of the graphical representation, the characteristics of the user and the nature of the task. The previous experiments explored the role of graphical features in comprehension performance.

Experiment 5

The aim of this study is to determine how, given the same open-ended task (try to understand what the graph is portraying), the nature of the interaction can also significantly affect performance. Specifically, I seek to determine whether the reduction in performance found in novice line graph users may be partially accounted for by the additional cognitive demands imposed by producing a think aloud protocol and whether this may be mitigated by engaging in a different way.

Method

Participants

Sixty-five undergraduate psychology students (54 female, 11 male) from the University of Huddersfield were paid £5 in vouchers to take part in the experiment. The age of participants ranged from 18.5 to 39.5 years with a mean of 21.5 years (SD = 3.8). All participants were in their first year of a three-year psychology degree.

Design

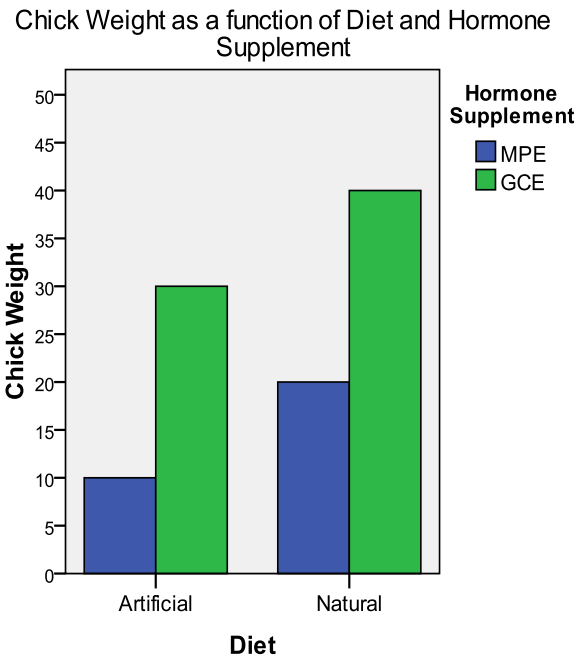
The experiment was an independent groups design with two between-subject variables: type of diagram used (bar or line graph) and nature of interaction with the graphs (think aloud or written responses). Sixty-five participants were randomly allocated to each condition. In the written condition booklets were mixed (so that the bar and line conditions were randomly mixed) and handed out to participants. In the think aloud condition participants were alternately allocated following the same procedure as experiment one. There were 14 participants in the verbal protocol bar condition and 16 in the written bar condition, 15 in the verbal protocol line condition and 20 in the written line condition.

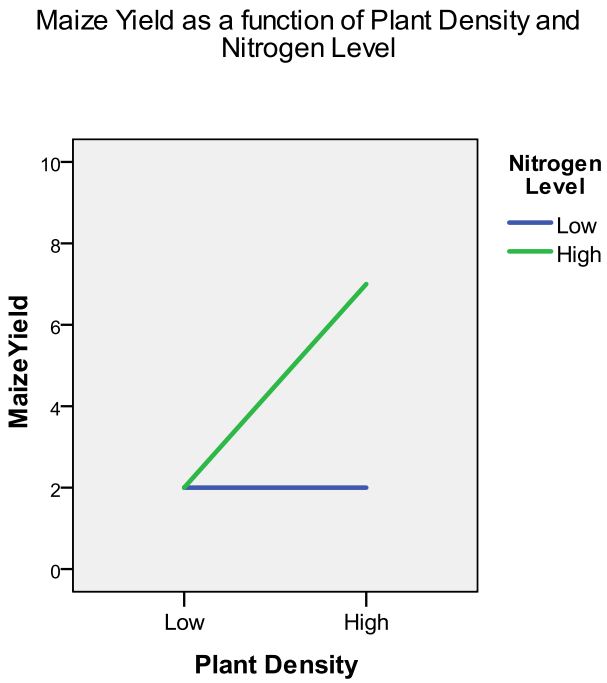
Materials

The stimuli used were six bar and six line three-variable interaction graphs depicting a wide range of (fictional) content. The graphs were generated using the PASW Statistics software package (produced by SPSS Inc.). Stimuli were printed in colour (with the levels of legend variable in blue and green) on white A4-sized paper.

Consistent with the previous experiments, the numerical values for the variables were selected in order to provide the range of effects, interactions and other relationships between three variables commonly encountered in these designs (typically depicted in line graphs as parallel, crossed and converging lines, one horizontal line and one sloped line, two lines sloping at different angles, etc.)

However, the content of these graphs differed to those used in experiments up to this point (see figure 17 for examples). Graphs depicting different content were introduced to demonstrate the findings were not specific to the relationships depicted in the stimuli used up until this point.





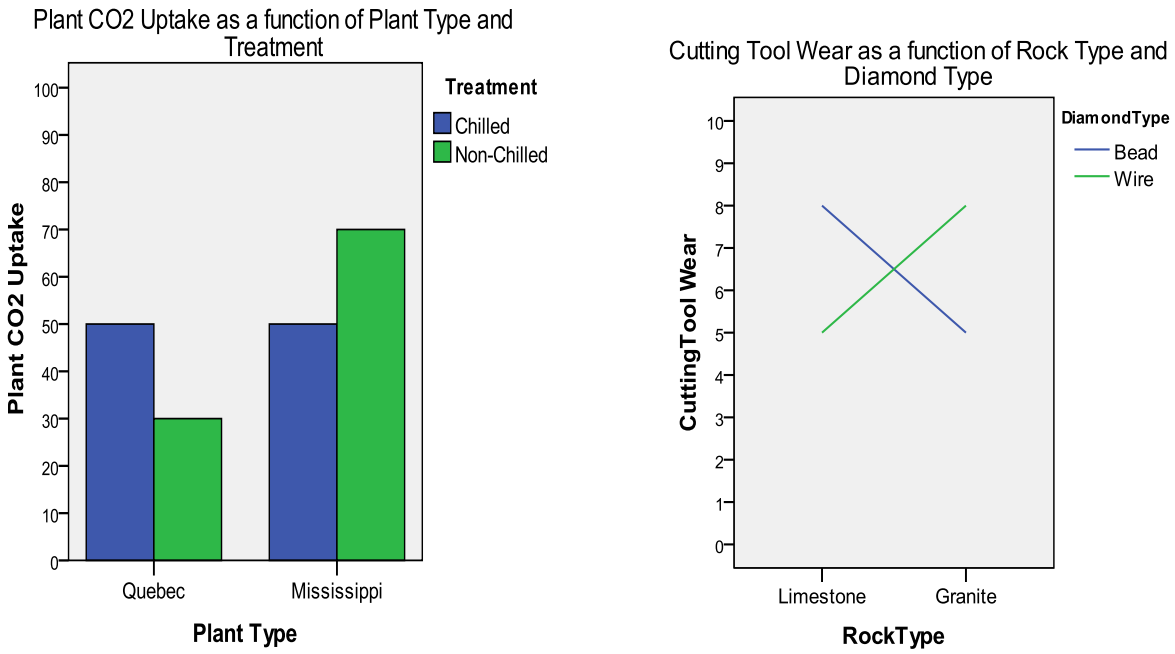


Figure 17: Bar and Line graphs representing the six data sets used in Experiments 5-7

Procedure

Participants were instructed that they would see six graphs and that their task was to try to understand each graph as fully as possible whilst writing their response down or thinking aloud. They were instructed to write or talk aloud about the relationships each graph was showing until they felt they had provided as much detail as they could.

The instructions drew attention to the fact that the graphs may depict more than one relationship and that participants should imagine they are in an exam in which more detailed interpretations produced higher scores. In order to produce as close a similarity as possible to the think aloud condition, participants in the written condition were also encouraged to write down their thoughts as they went along.

In the written condition the six stimuli were compiled as a booklet with graph pages interleaved with blank paper for writing. Participants completed these under the supervision of the experimenter. In the verbal condition the graphs were handed over to participants one at a time for them to interpret while their verbal

protocols were recorded using a portable digital audio recorder. Stimuli were presented in random order and all participants were informed that there was no time limit to the task.

Results

The analysis employed earlier to categorize participants graph reading ability was employed again for this experiment. The pattern of pre-elementary performance in the think aloud condition was consistent with previous experiments, a substantially higher proportion of line graph users were classified pre-elementary compared to bar graph users. This effect has consistently been demonstrated in numerous experiments where the proportion of graph users in each category are consistent across experiments employing first and second year undergraduate psychology students and final year students close to graduating.

Furthermore, the findings from this experiment demonstrate that the same pattern of results emerge when graphs depicting different content are used as the stimuli. Figure 19 displays the number of users in each category of graph reading ability.

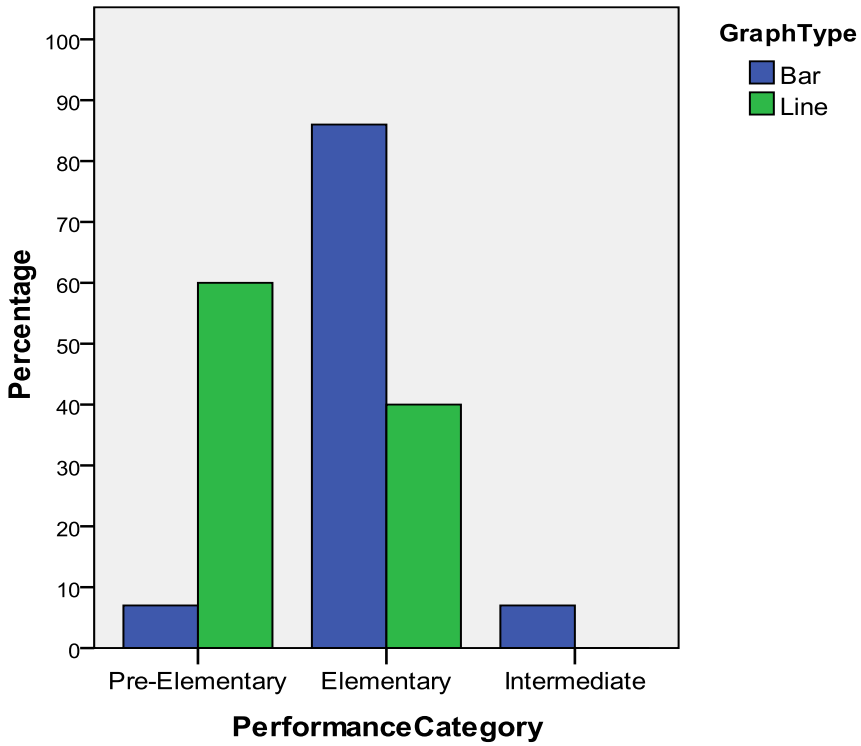


Figure 18: Percentage of bar and line graph users in the three performance categories, verbal protocol condition.

The bar line difference in the think aloud condition emerged as predicted; 60% of participants were classified as pre-elementary in the line graph condition compared to 7% in the bar graph condition. A chi-squared test of independence revealed that this association between line graph users and pre-elementary performance was statistically significant (chi-square = .82; df = 1; $p < .01$), replicating the result of the original Peebles and Ali (2009) experiment and the experiments reported earlier (Experiment 1 and 4). Consistent with the previous analyses, I also analyzed the data by trial. This revealed that the mean ranks (11.0) in the think aloud line graph condition was significantly lower than in the bar graph condition (19.29) $U = 45$, $p < .01$.

However, a different pattern of performance emerged in the written condition. Analysis of written responses revealed a remarkable drop in pre-elementary performance in the line graph condition. In the written condition the high rate of pre-elementary performance found in the think aloud line graph condition was not replicated, with a considerably lower percentage of participants classified as pre-elementary in the written line graph condition (15%). The number of participants classified as pre-elementary was roughly equal between the two graph formats in this condition with 19% of participants classified as pre-elementary in the bar graph condition. Figure 18 displays the number of users in each category of graph reading ability.

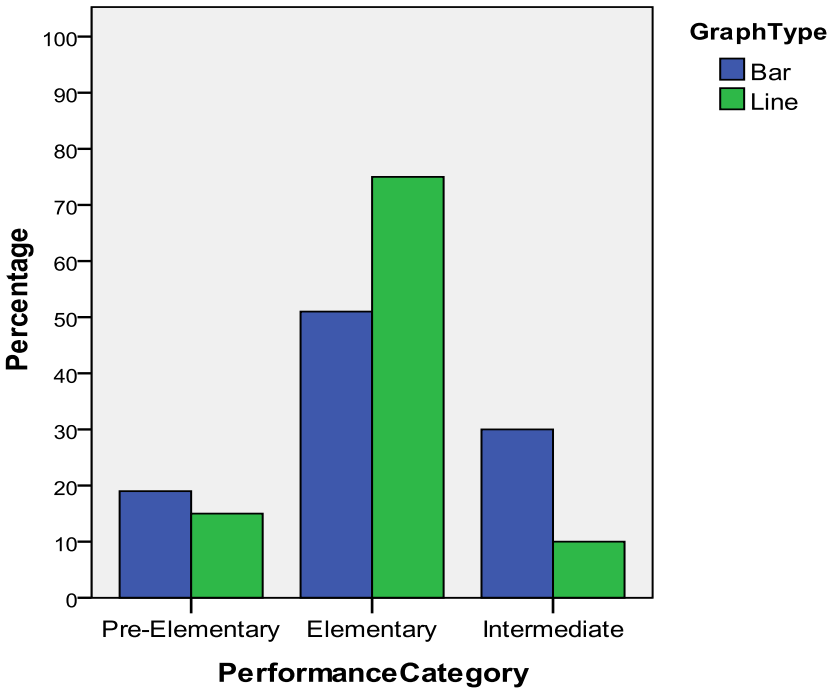


Figure 19: Percentage of bar and line graph users in the three performance categories, written condition.

Similar to the colour match graph the majority of participants were classified as elementary and a smaller proportion pre-elementary and intermediate. The proportion of users classified as intermediate was slightly higher in the bar graph condition than the line graph condition.

Requiring participants to write their answers down resulted in a number of improvements across different measures. Firstly, viewers' understanding of relationships depicted in graphs improved dramatically. Perhaps the most striking finding was that 45% of participants interpreted all six trials correctly in the line graph condition and 50% in the bar graph condition. This compares favourably to 13% of line graph readers (29% of bar graph readers) in the think aloud condition. The improvement that emerged from writing was very similar to the improvements observed in the colour match graph. However, a Fisher's Exact test found that the association between the number of participants interpreting all six trials correctly in the written line and think aloud line graph condition was not significant ($p = .07$) nor was the association between the written bar or think aloud bar graph condition (chi-square = .14, df = 1 $p = .23$).

A Fisher's Exact test revealed that the number of participants classified as pre-elementary between the two graph formats in the written condition was not significant (chi-square = .09; df = 1; $p = 1.0$). In the written condition mean ranks were similar (bar = 19.41, line = 18.69). There was no significant difference in number of correct trials between the two graph conditions ($U = 161.5$, $p = .84$).

A comparison of the number of correct trials for the written bar and verbal bar condition revealed they were similar (written mean ranks = 16.16, verbal mean ranks = 14.75). This difference was not significant: $U = 141.5$, $p = .67$.

However, there was an interaction effect of graph format and nature of interaction; participants conceptual understanding of the line graphs was superior in the written line condition (mean ranks = 22.71) than the think aloud condition (mean ranks = 12.60) $U = 69.0$, $P < .01$.

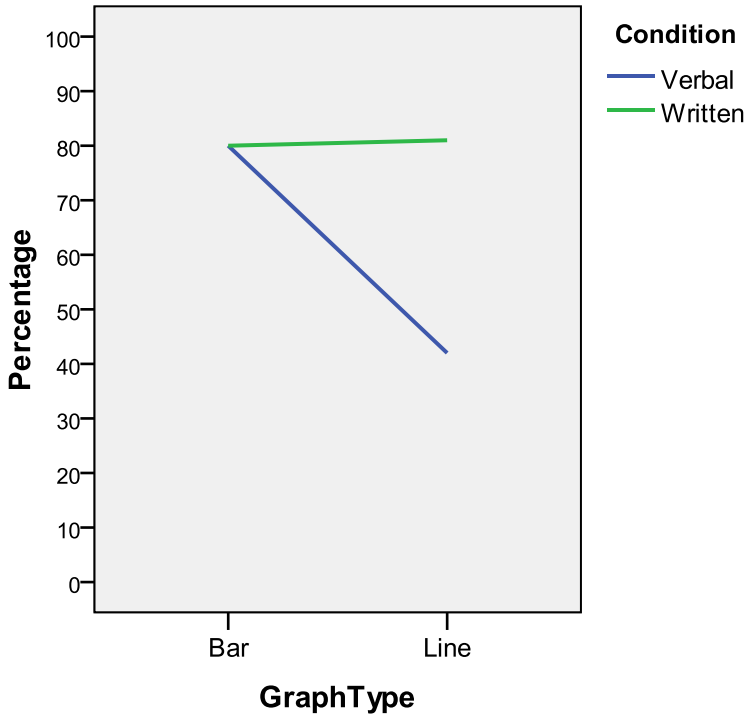


Figure 20: Percentage of correct trials for bar and line graphs in the verbal protocol and written conditions

Figure 20 reveals how the number of correctly interpreted trials was similar in the bar graph condition irrespective of nature of interaction with the graphs. However, there is a marked difference in graph readers' ability to provide a correct interpretation depending on the way in which an interpretation is provided. Those required to think aloud whilst interpreting the graphs were substantially less likely to interpret graphs correctly than those required to provide a written response.

Discussion

The results of this experiment reveal a remarkable interaction between graph format and the type of interaction with the diagram. Consistent with the results of earlier experiments and those reported in Peebles and Ali (2009), a significant proportion of line graph users were classified as pre-elementary compared to bar graph viewers in the think aloud condition.

However, this effect does not emerge in the written response condition. Despite the imbalance of gestalt principles associating the pattern to referents, the majority of graph readers demonstrate conceptual understanding of both graph formats at an elementary - and in a few cases intermediate – level. Results reveal differences in conceptual understanding of graph formats are due to the type of interaction with the diagram. These results clearly demonstrate written responses are superior to verbal protocols for what students find the more difficult graph format – line graphs (Friel, Curcio and Bright, 2001).

There are a number of potential competing explanations for why this difference in conceptual understanding in the line graph condition emerges for these two different types of interaction. Written responses and the verbal protocol method vary in a number of ways.

One key difference between the two conditions is presence of the experimenter – although the experimenter was present in both conditions (written and verbal protocol), in the latter the participant may be more acutely aware of the experimenter's presence as they are having to verbalize their thoughts to them and so may feel pressured to present themselves in the best possible light. As the majority of line graph users struggled to understand the graphs they were presented with presence of an observer could have resulted in a detriment in performance.

Therefore, the social nature of this interaction could be impairing performance in the line graph condition. This potential explanation gains some support from participants' reactions to undertaking the task. The majority of participants expressed negative self-evaluations during or after the task. For example, participants would frequently say "I sound thick!", "you must think I'm really stupid", "this shouldn't be so hard". In the written condition the experimenter is removed from the participants focus as they are not

required to verbalize their thoughts. Therefore, any difficulty they may be experiencing is not externalised and so there is no performance evaluation occurring. This difference could potentially explain the difference in participants' performance.

Secondly, the think aloud method requires participants to verbalize their thoughts throughout the task whereas in the written condition these demands are not present. It is possible the demands posed by verbalization are interfering with the task and resulting in reactivity effects – in this case a detriment in performance in the line graph condition. Some support for this hypothesis comes from literature investigating whether the act of thinking aloud alters underlying cognitive processes involved in the task being undertaken (Russo, Johnson and Stephens 1989).

Finally, it is possible that the task demands differ between these two types of interaction. In the written condition it is clear that the writer has to communicate their understanding to someone else – participants wrote in booklets which were being returned back to the experimenter. In the think aloud condition participants' primary focus may be on understanding the data for themselves and so the task requirement of explicitly communicating understanding to someone else may be inhibited by the demands to think aloud.

This experiment does not allow me to differentiate between these competing explanations. The following experiments will attempt to isolate which factor is causing the difference observed between conditions. Only the line graph condition will be tested, as conceptual understanding of bar graphs is broadly consistent across different types of interaction. The question of interest concerns why a dramatic difference in conceptual understanding of line graphs is found when different types of interaction are employed to assess understanding of these diagrams.

Chapter 8

Why does the verbal protocol method result in a high rate of pre-elementary performance in the line graph condition?

Although the verbal protocol method has come under a considerable deal of scrutiny as a valid process tracing method, there has been little attention paid to the question of whether experimenter presence impacts upon the type of protocol participants provide. This is probably because experimenters are present during data collection for a number of research methods and experimenter effects are broadly covered in literature concerning advantages and disadvantages of research methods.

Reviews of the literature have revealed numerous factors which could possibly influence results, ranging from experimenter's sex, authoritarianism, birth order, intelligence, age, race, religious background and anxiety level. As well as reviewing literature investigating attributes of the experimenter, research measuring participant attributes and interactions between experimenter/participant attributes has been analyzed (Rosenthal, 1976).

This research has demonstrated that the gender of the experimenter and participant can influence results of research. Stevenson and Allen (1964) employed both male and female experimenters. Participants were required to classify marbles by colour. Both male and female experimenters praised participants on their performance. They found a significant interaction between gender of experimenter and gender of participants. When male participants were paired with a female experimenter and female participants paired with a male experimenter, a significantly higher number of marbles were catalogued than when the researcher and participant were of the same gender.

In an attempt to explain their results the authors suggested that the observed interaction could be because of increased competitiveness, increased anxiety or greater desire to please when the experimenter is of a different gender. However, they also accepted it could be because experimenters treat members of the opposite sex differently to those participants who are the same sex as them.

Research has shown that experimenters behave differently towards male and female participants in an experimental situation. Female participants receive more eye contact from experimenters than male participants as the experimenter glances at them a lot more frequently than male participants (Freidman, 1964, Katz, 1964, cited by Rosenthal, 1976). This led Rosenthal (1976) to conclude that female participants are treated more courteously than male participants. Furthermore, there are differences resulting from the gender of the experimenter as well. Findings show female experimenters smile more often at participants than male participants (Katz, 1964).

When measuring the number of verbal responses participants could come up with, Reece and Whitman (1962) found that the "warmth" of the experimenter affected results. They defined "warm" behaviour as leaning in the direction of the participant, glancing at them, smiling and keeping hands motionless. Conversely, "cold" behaviour was operationalised as leaning away from the participant, looking around the room, not smiling and drumming their fingers. Their results revealed that verbal output was larger when participants were in the "warm" condition.

Rosenthal (1976) also considered characteristics of participants which make them want to please the experimenter and want to participate "correctly" so that they provide the results the experimenter wants. He highlights a study where after completing the experiment one participant asked "did I do it right?" Another participant was explicit about their worries and queried "I was wondering if I was doing the experiment the way it should be done".

This line of research clearly demonstrates that there are numerous variables which may influence participants' performance during an experiment including their own personal attributes such as the need to please. Researchers who are present or interacting with participants throughout an experiment need to be aware of such variables, although it would be impossible to control unconscious behaviour being emitted by the experimenter. This line of research is important for researchers employing the verbal protocol method as the researcher is present throughout the experiment, although every attempt is made to minimize presence and interaction with the experimenter during the experiment.

This issue has been specifically brought to light with this research because it is not always possible to avoid interaction with participants during the experiment. Typically, the large majority of participants who took part in this research study found the task difficult (this was especially pronounced in the line graph condition) and often asked the experimenter for reassurance or guidance. For example, participants would often say "is that right?" or "is that what I was supposed to say?" or as pointed out earlier express negative self evaluations such as "I sound so stupid."

This was despite the fact that instructions emphasized the experiment was not a test of their ability. Participants even exhibited signs of stress with some rubbing their heads, or more commonly laughing nervously. Some participants even apologized for "messing up the results". This again was despite assurances that there was no single correct response. It would seem in tasks in which participants are required to interpret material but struggle to understand the material they will automatically assume the experimenter is looking for "good" performance in terms of them providing a sophisticated interpretation.

Perhaps the verbal protocol method is not an appropriate methodology to employ if participants are struggling with the task. Although experimenter effects have only been investigated broadly and not specifically in relation to the verbal protocol method there is research to suggest that presence of others affects performance in either a positive or negative way. This field of research comes under the umbrella of social facilitation/inhibition.

Initially, the field of social facilitation emerged based on findings that both people and animals perform better when in the presence of others than when alone (Zajonc, 1965). This finding was termed "social facilitation", to describe the enhancing effects of the presence of others on performance. However, this effect was not consistent and further research revealed that presence of others resulted in poorer performance than when working alone. This led to the term "social inhibition" to describe the inhibitory effects of presence of others on performance. In an attempt to explain these inconsistent findings Zajonc (1965) proposed a drive theory of motivation which predicts that the presence of others enhances performance on simple or well learned tasks but inhibits performance on difficult or unfamiliar tasks. A meta-analysis of 241 experiments confirmed this hypothesis (Bond and Titus, 1983).

This theory proposes that a high level of arousal boost the dominant response. Zajonc (1965) argued that presence of others increases an individual's level of arousal, which in turn aids the dominant response. In the case of easy or well learned tasks the dominant response would be the right answer and so an improvement in performance results. Conversely, for difficult or novel tasks the dominant response would probably be incorrect and so performance deteriorates. Experimental research has confirmed this hypothesis. For example, participants learn easy words more quickly in the presence of an audience but lean difficult words more slowly (Cottrell, Wack, Sekerak and Rittle, 1968)

An alternative explanation for this pattern of findings is concerned with attention. Fundamental to these theories is that presence of others is distracting, which results in a more constricted focus of attention (Baron, Moore and Sanders, 1978). This theory can also explain the social facilitation / inhibition effect, as performance should be better when participants need only focus on a small number of cues but inhibition should occur when attention is required for a large number of cues (difficult tasks).

These two competing theories make the same predictions concerning task performance, making it difficult to distinguish which provides a more adequate explanation of how the presence of others affects task performance. However, a study by Huguet, Galvaing, Monteil & Dumas (1999) used the stroop task, where each theory would make differing predictions. The stroop task is a difficult untaught task which contains only a few stimuli. This task requires individuals to identify the colour words or symbols are printed in. A robust finding concerning this task is that individuals can identify the colour of symbols quickly but slow down when the stimuli consists of words that are not consistent with the ink colour (for example the word yellow printed in red ink).

This effect occurs because reading is a dominant and automatic response in adults, so when they are instructed to name the ink colour and ignore the written word interference occurs. In relation to competing predictions of drive theories and attention theories, the drive theory would predict poor performance as word reading is such a dominant response that interference will occur because the dominant response has been heightened due to the presence of an experimenter.

However, because the stroop task contains only two stimuli (the word and ink colour) and constricted attention will lessen attention to extraneous stimuli (the word) attention theories would hypothesise that presence of others would enhance performance and so social facilitation should occur. In an investigation of which theory could better account for the results in the area of social facilitation research Huguet et al (1999) employed the stroop task. They found predictions from attention theories were supported by individuals' performance on the stroop task; people performed better when others were present.

Another competing theory attempting to explain why presence of others results in a change in performance is evaluation apprehension approaches. These research studies suggest that the experimenter could be perceived as evaluative. In a review of studies into how presence of others impacts performance Guerin (1986) found that from 39 of the experiments included for review, 34 found experimenter presence effects.

One of these studies found that participants viewed the experimenter as being an expert which in turn could lead to evaluation apprehension. Scotland and Zander (1958, cited by Guerin, 1986), conducted an experiment where participants worked in the company of either a researcher who stated they were a professional in the area or one who stated they did not know much about the area. They found that the experimenter claiming to be a professional was evaluated by participants as being more knowledgeable but the researcher who claimed to know little about the area was still considered as being relatively expert. This suggests that a researcher can be considered to be an expert because they are present.

There are competing theories of why evaluation apprehension would result in detriment in performance of complex tasks. Generally researchers argue that it is due to evaluation apprehension from being observed, but similar approaches (Bond, 1982) suggest that it could be due to self-presentation effects (effects occur because individuals wish to maintain a certain public image). These approaches are not incompatible – they can be seen as different features of the same effect – attempts by research participants to gain and sustain public approval (Guerin, 1986).

Therefore, the next experiment investigates whether the presence of the experimenter inhibits performance in the line graph condition. It may be the case that participants are experiencing evaluation apprehension because they are aware that the experimenter is listening to them think aloud.

Experiment 6

The aim of this experiment is to determine whether experimenter presence has a detrimental effect in the line graph condition. In order to investigate this research question participants were left alone whilst completing the experiment – and it was made clear they would be alone throughout and not interrupted. If experimenter presence is the underlying cause of poorer performance in the line graph condition then there should be an improvement in performance when participants are left alone to do the task.

Method

Participants

Fifteen undergraduate Psychology students (9 female, 6 male) from the University of Huddersfield volunteered to take part in the experiment for which they were paid £5 in vouchers. The age of participants ranged from 18.9 to 24.6 years with a mean of 19.9 years (SD = 1.8). All participants were in the first year of a three year Psychology degree.

Materials, Design and Procedure

The experiment was carried out using the same equipment and the same procedure as Experiment 5 (verbal protocol condition), the only difference being that there was only one graph condition in this experiment and participants were left alone during the experiment whilst completing the task.

Results

Results were consistent with previous findings – a high proportion of line graph users were classified as pre-elementary in this condition (see figure 21). No participants were classified as intermediate and identical to the think aloud condition only 13% of the sample interpreted all six trials correctly.

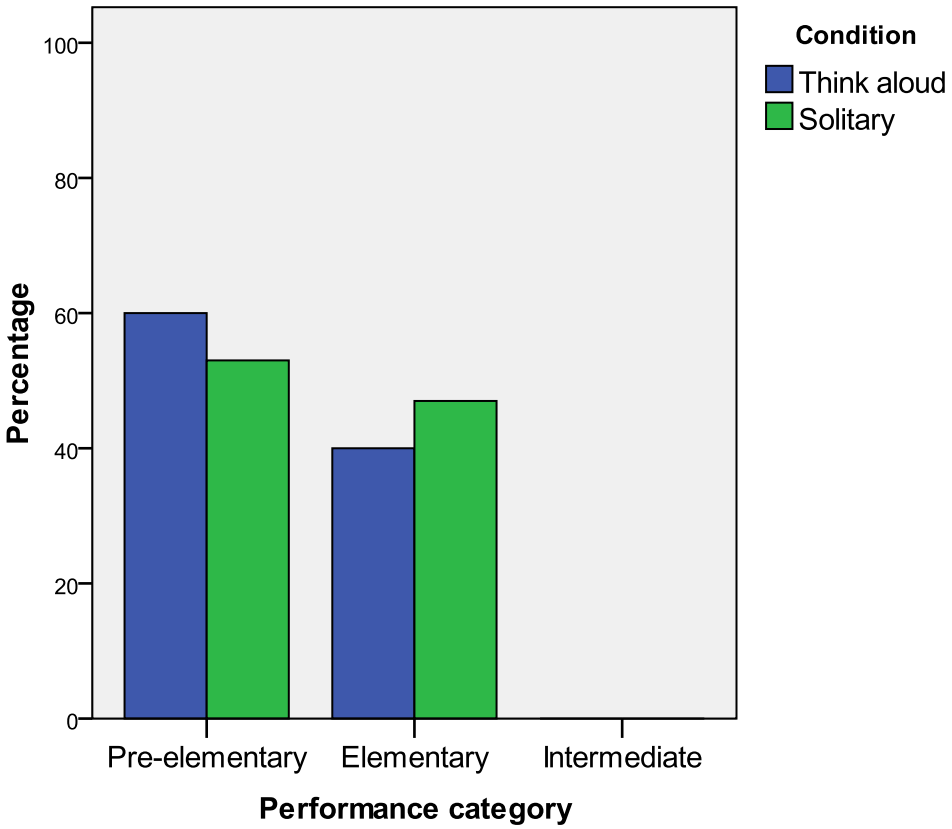


Figure 21: Percentage of line graph users in the three performance categories, think aloud and solitary condition

53% of the sample was categorised as pre-elementary in this condition. A chi-square test of independence compared the number of pre-elementary users in this experiment to the think aloud line graph condition in Experiment 5 and found the association between the two conditions was not significant (chi-square = .71, df = 1, p = 1.0). An analysis of the number of correct trials revealed participants performed slightly better in this condition (solitary mean ranks = 14.9, think aloud mean ranks = 16.10), this difference was not significant (U = 103.5, p = .70).

Discussion

The high rate of pre-elementary performance observed in the earlier experiments where participants were required to think aloud with the experimenter present was replicated in this condition where the experimenter was absent. This suggests that it was not experimenter effects or some form of social inhibition resulting in the poor performance in the line graph condition.

Despite the established effect in the literature demonstrating superior performance in simple tasks but poorer performance in complex tasks in the presence of others, this effect does not occur for these particular tasks. This may perhaps be because these tasks are knowledge based tasks whereas literature focussing on how the presence of others affects performance concentrates on tasks that do not require specific knowledge. For example, paired associate word tasks (Geen, 1983) or dressing or undressing in familiar or unfamiliar clothing (Markus, 1978).

Evidence for this is provided by the studies conducted by Schooler, Ohlsson and Brooks (1993) who discounted a social inhibition explanation for the pattern of results found in their experiments. They found participants performed significantly worse when solving insight problems than non-insight problems whilst verbalizing their thoughts concurrently. They matched insight and non-insight problems so they were equally difficult. Their findings, that verbalization impaired problem solving for insight but not non-insight problems ruled out a social inhibition effect explanation, because if this was the explanation a detriment in performance should have occurred for the non-insight problems as well.

As experimenter presence has been ruled out as a potential explanation for the differing results in the written and think-aloud condition, the next major difference between the two conditions will be tested. The results of the research conducted by Schooler, Ohlsson and Brooks (1993) and Russo, Johnson & Stephens (1989) indicate the requirements to verbalize may result in reactivity effects for certain types of problems. Therefore, the next experiment will investigate whether performance differs when participants do not need to verbalize their thoughts throughout the task.

Experiment 7a: Do demands of verbalization interfere with task demands?

An obvious way in which the verbal protocol method differs from the written method is the demands to verbalize in the verbal protocol condition. Employing Type 1 verbalizations which require participants to concurrently think aloud whilst carrying out the task could potentially add additional demands not present in the written condition.

The possibility of verbalization interfering with the primary task is known as “reactivity effects” (Russo, Johnson & Stephens, 1989) and whether the act of producing a protocol is reactive is typically investigated

by incorporating a silent condition in the experiment. In one condition participants provide a protocol whilst completing the task and in another condition complete the same task silently. Output measures are recorded (for e.g., response time, number of correct responses) and compared to performance in the think aloud condition (Ericsson and Simon, 1993).

Therefore, in order to test whether it was providing a protocol that resulted in a detriment in performance in the think aloud condition a silent condition was included as an experiment. In this condition participants were required to interpret the graphs but at first they spent time reading the graphs silently. Once they felt they had understood the graphs as much as possible they then verbalized their interpretation to the experimenter.

The aim of this experiment is to determine whether remaining silent whilst engaging in the task resulted in any discernible benefits.

Method

Participants

Fifteen undergraduate Psychology students (11 female, 4 male) from the University of Huddersfield volunteered to take part in the experiment, for which they were paid £5 in vouchers. The age of participants ranged from 18.1 to 23.2 years with a mean of 20.5 years (SD = 1.5). All participants were in the first year of a three year Psychology degree.

Materials, Design and Procedure

The experiment was carried out using the same equipment and the same procedure as Experiment 5 (verbal protocol condition). Participants were instructed that the experiment consisted of two stages – in the first "quiet" stage they could take as long as they wanted to understand the graph they were viewing as much as possible. In the second "talking" stage they were required to tell the experimenter what they had understood about the graph.

Results

Results were similar to the written condition – only a small proportion of the sample was classified pre-elementary. The high rate of pre-elementary performance found in the think aloud condition in Experiment 5 was not replicated with only 7% of the sample classified as such in this condition. Similar to the colour match graph and written interpretations, the majority of participants were classified as elementary and a smaller proportion pre-elementary and intermediate. Figure 22 displays the number of users in each comprehension category, alongside the results of the think aloud condition in Experiment 5 for comparison.

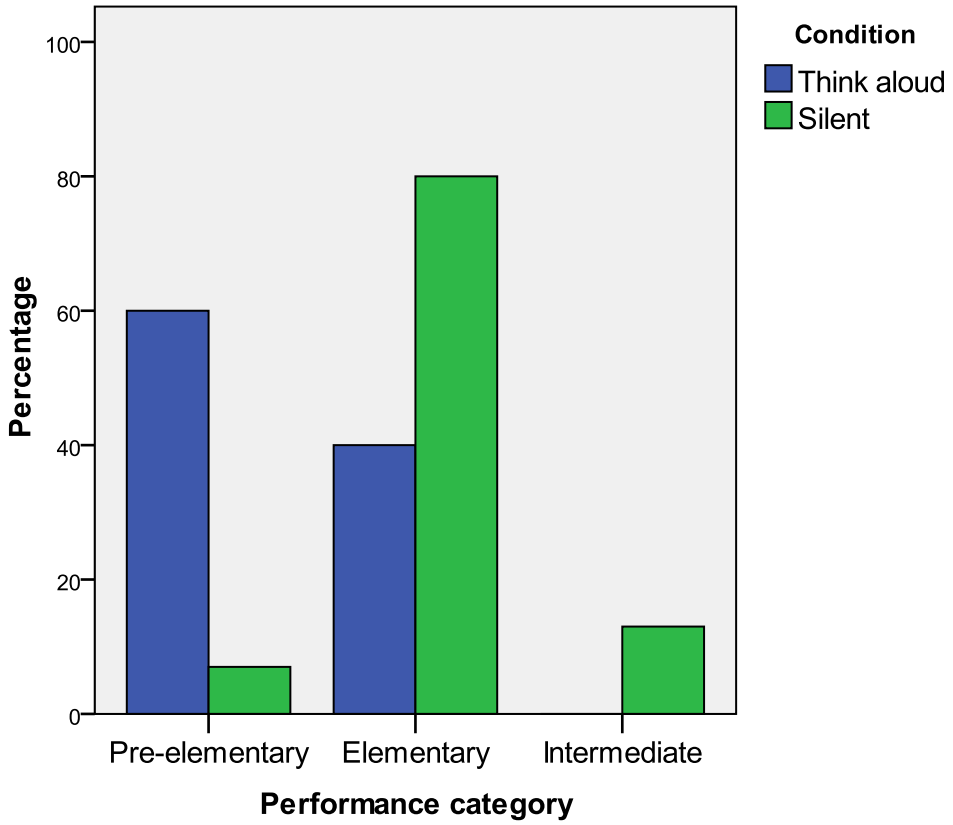


Figure 22: Percentage of line graph users in the three performance categories, think aloud and silent condition

The improvement that emerged from the silent condition was very similar to the improvements observed in the written condition and the colour match graph. 40% of participants interpreted all six trials correctly in this condition. This compares favourably to 13% of line graph readers in the think aloud condition. However, this association was not significant (Fisher's Exact test $p = .22$).

A chi-square test revealed that the number of participants classified as pre-elementary in the silent and think aloud line graph condition was significant (chi-square = 10.4; df = 2; $p < .01$). A comparison of the number of correct trials between the conditions also revealed that the silent condition resulted in a significant increase in the number of correctly interpreted trials (think aloud mean ranks = 11.7, silent mean ranks = 19.93). This difference was significant: U = 46, p = <.01

Discussion

The results of this experiment suggest that the high rate of pre-elementary performance consistently found in the line graph condition in the think aloud conditions was a result of the demands of verbalizations. When participants were allowed to remain silent whilst formulating their interpretation pre-elementary performance was equal to the bar graph condition and level of comprehension was similar to the written condition.

Therefore, despite the impressive research evidence reviewed by Ericsson and Simon (1993) supporting their assertion that protocols are not reactive it would appear that the conclusion drawn by Russo, Johnson & Stephens (1989) receives further support from the results of this experiment – the effect of providing a protocol is dependent on the interaction between task demands and the demands of verbalization.

There are only a limited number of studies in the literature which have found reactivity effects in tasks employing the think aloud method. In fact, Russo, Johnson & Stephens (1989) concluded that because of the lack of studies finding reactivity effects, there is an agreement in the literature that employing the think aloud method slows down processing but does not alter it.

To determine whether providing a protocol can be reactive Russo, Johnson & Stephens (1989) tested four different tasks in an attempt to test Ericsson and Simon's (1993) claim that verbal protocols are accurate reflections of underlying cognitive processes. They found that for two of the four tasks - Raven's Matrices and anagrams - providing a protocol resulted in no differences between the silent and think aloud condition.

However, for the other two tasks the effect of thinking aloud differed depending on the nature of the task. Thinking aloud whilst making a choice between two gambles significantly improved task performance compared to the silent condition. However, when participants were required to add three digit numbers

performance deteriorated. This led them to conclude that the reason why reactivity effects did not often occur is because the demands of verbalization interact with the demands of the task to affect output.

Russo, Johnson & Stephens (1989) proposed a processing resources explanation to account for their results. They suggest that when the demands of verbalization compete with task demands then participants will have to choose where to allocate processing resources. Russo, Johnson & Stephens (1989), following on from Kahneman (1973) suggest that participants draw on slack resources which are not being used up by the task to verbalize whilst completing the task. If the demands to think aloud are minimal and resources are available then there will be no change in task performance.

However, when the availability of slack resources is minimal and demands to verbalize requires more resources than are available then participants face a choice: they can either continue to verbalize which will result in a drop in the resources available for the task or stop thinking aloud which violates the instructions of the task. If participants choose the former option then reactivity effects can result.

Therefore, when participants are required to solve a difficult problem then there will be little or no processing resources available for the demands posed by verbalization, whereas if the problem is simple then there are slack resources available to draw upon, thus resulting in little or no difference in problem solving despite the demands to verbalize. Kahneman (1973) suggests the more difficult the problem the higher the likelihood problem solving will be impaired if there is a competition for resources.

This explanation could account for the pattern of results found for the experiments conducted for this research. Although it could be problematic defining line graphs as more difficult to interpret than bar graphs, Kahneman (1973) suggests problem difficulty could be measured by number of errors made whilst solving problems. Since the pattern of results from these experiments consistently show participants make more errors in the line than the bar graph condition, the line graph format can be considered to be more difficult to interpret than bar graphs.

Russo, Johnson & Stephens' (1989) finding that task demands interact with the think aloud method to affect output is consistent with the findings of this research. Performance in the bar graph condition appears to be

unaffected by the demands to think aloud whereas the requirement to think aloud in the line graph condition results in a marked detriment in performance when compared to a silent condition.

Therefore, one potential explanation for these findings would be that when interpreting line graphs - which are difficult to understand without training - there is a high degree of competition for processing resources. When additional demands posed by verbalization are also present then resources required to interpret the information presented in line graphs are diverted to demands posed by the instructions to verbalize. This competition for resources could potentially result in the pre-elementary performance observed in the line graph condition in numerous experiments which does not emerge when the demands to verbalize are not present – in the silent and written condition.

However, this proposed explanation by Russo, Johnson & Stephens (1989) has been challenged by Schooler, Ohlsson and Brooks (1993) who proposed a very different explanation for why reactivity effects occur with certain types of problems. Schooler, Ohlsson and Brooks (1993) tested for reactivity effects by comparing performance on insight and non-insight problems whilst participants completed the problem silently or whilst thinking aloud. They found that the demands of thinking aloud resulted in significantly fewer insight problems being solved compared to non-insight problems.

As these two types of problems were matched for difficulty, Schooler, Ohlsson and Brooks (1993) argued that the proposed explanation by Russo, Johnson & Stephens (1989) could not explain their results. If a processing resources explanation was valid then a detriment in performance should occur in the non-insight problems where participants were also required to think aloud throughout. However, they found no detriment in performance for these types of problems. Therefore, they concluded that verbalization was not reducing the availability of resources to solve a problem.

Instead, they explain their results by drawing on the verbal overshadowing paradigm. They argue that the demands to verbalize directs the way resources are allocated, rather than consuming resources and making them unavailable for use in the primary task. Specifically, attention is directed to aspects of the problem that are easy to verbalize which draws attention away from processes which are difficult to verbalize.

This explanation could also perhaps explain the results of this research. An analysis of the Gestalt principles present in interaction bar and line graphs revealed that bar graph displays have a perceptual feature allowing readers to relate the pattern to both independent variables. However, the line graph display has a perceptual feature allowing readers to relate the pattern to the legend but not to variables plotted on the x axis.

Therefore, the verbal overshadowing explanation proposed by Schooler, Ohlsson and Brooks (1993) could be applied to these types of diagrams. Although not insight problems it is possible that the demands to verbalize draws attention to information that is easy to verbalize in the line graph condition. This information would be the variables in the legend, as a simple colour matching process allows participants to match the lines at the centre of the display to the variables in the legend.

As there is no equivalent grouping process available allowing readers to relate the pattern to variables plotted on the x axis, this information would be difficult to verbalize and so attention would be drawn away from it. This could perhaps explain the pattern of errors found when participants are required to interpret line graphs whilst thinking aloud – the large majority of participants ignored the x variable and described the relationship between the legend and y axis variable.

Although Schooler, Ohlsson and Brooks (1993) discount a processing resources explanation of reactivity effects and present convincing evidence this theory cannot account for their results, this explanation cannot be discounted completely. This is because the task Russo, Johnson & Stephens (1989) found a significant detriment in performance in does not consist of information that is difficult to verbalize. Their task required participants to add three-digit numbers, information which is consistent with a verbal code and so easy to verbalize (Ericsson and Simon, 1993).

The research literature demonstrating reactivity effects which occur from thinking aloud whilst completing a task is limited and therefore any theoretical account attempting to explain such effects is at an early stage. It is possible that the conclusions Russo, Johnson & Stephens (1989) drew about thinking aloud interacting with task demands to affect the output of a task can also be applied to the two theoretical accounts proposed to explain reactivity effects.

It is possible that when information is difficult to verbalize, attention is focussed on the components of a problem that are easily to verbalize (Schooler, Ohlsson and Brooks, 1993). This does not necessarily discount the explanation however, that processing resources are divided between task demands and the demand to verbalize which can cause a detriment in performance (Kahneman, 1973 Russo, Johnson & Stephens, 1989).

Based on the results of this experiment it would appear that the pre-elementary performance found in the think aloud line graph conditions is a result of the requirements to verbalize. However, due to the nature of the task it could be argued that the second stage of the task – interpreting the graphs out loud to the experimenter, could be influencing the results. The communicative aspect of providing an interpretation to someone else could be resulting in an improvement. Therefore, splitting the task into two stages could potentially result in a confounding variable. Comparing the silent condition to the think aloud condition is not appropriate because of this potential confound.

Therefore, the final experiment will test the explanation that communicating understanding to someone else is resulting in an improvement in performance. Again the task will be split into two stages. In the next experiment participants were required to interpret the graphs whilst thinking aloud so the first stage of the task was identical to the think aloud condition. Once they felt they had understood the graphs as much as possible, they then provided an interpretation to the experimenter.

Experiment 7B: Does communicating understanding to someone else improve conceptual understanding?

The previous experiment examined whether allowing participants to remain silent before providing an interpretation improved conceptual understanding of line graphs. A significant improvement was found indicating the requirement to think aloud whilst undergoing a task is detrimental to understanding of material. However, the task also required participants to provide an interpretation to the experimenter. Being explicitly required to communicate understanding to someone else could perhaps result in a performance improvement that is unrelated to undergoing a task silently.

The notion that communicating understanding can in itself improve comprehension of material has been investigated. This research comes under the umbrella of "self-explanation" where findings have revealed that requesting participants to explain material they are engaging with improves understanding. Findings reveal the more participants self-explain, the higher the rate of success when problem solving or demonstrating understanding of material (Nathan, Mertz, and Ryan, 1993, Pirolli and Recker, 1994).

Chi, Leeuw, Chiu and Lavancher (1994) investigated whether students' conceptual understanding of a biology topic - the human circulatory system – benefitted from the students being instructed to self-explain. Eighth grade students were instructed to read material from a biology textbook. One group was instructed to state what they had understood about the material whereas another group were instructed the read the same material twice. Knowledge gain was assessed by including a pre-test and post-test. Results revealed that those who self-explained showed greater gains in knowledge than those who had simply read the text twice. Consistent with previous research they also found that those students who provided more explanations showed greater gains from pre-test to post-test when compared to those individuals who provided fewer explanations.

This area of literature is consistent with the gains found when Type 3 verbalizations are employed to investigate a research question. In their analysis of how different types of verbalizations elicit different responses Ericsson and Simon (1993) reviewed studies which required participants to explain the choices they made during a task. A consistent finding which emerged from the review of research studies was that requiring participants to explain their decision making or thought process resulted in an improvement in subsequent performance.

One example is a study conducted by Ahlum-Heath and Di Vesta (1986) which investigated participants' ability to solve the Tower of Hanoi problem. They found that requiring participants to provide a reason for the move they made prior to the move being made facilitated performance when compared to the no verbalization group. Based on their results they concluded that although practice facilitated performance more so than no practice, it was the combination of practice with explanatory verbalizations which result in the highest performance improvements.

Similarly, Stanley, Matthews, Buss and Kotler-Cope (1989) developed specific verbalization instructions which they hypothesised would result in an improvement in performance in problem solving. They designed this new set of instructions because prior to them Berry and Broadbent (1984) found no improvement from instructing participants to verbalize concurrently using the same task. In the experiment conducted by Stanley, Matthews, Buss and Kotler-Cope (1989) participants were instructed to imagine they were managers of a sugar factory and their task was to attain and sustain a certain level of sugar production.

Instead of asking participants to think aloud Stanley et al (1989) told participants they needed to provide information to a partner they would not meet would then undergo the same task with the original participants' instructions to guide them. Participants were told: "Please give your instructions for your partner. Try and be as complete and specific as possible in telling him or her how you are making your choices. Try to give him or her more information than you did in your last instruction" (p.559). After each block of trials there was a pause to allow participants to give instructions so participants were not verbalizing and performing the task at the same time.

Stanley et al (1989) found that the condition where participants were providing verbal instructions significantly outperformed the control condition that underwent the task silently. As the only manipulation was the requirement to give instructions, the authors concluded that using a specific type of verbalization procedure can result in improvements in task performance. This verbalization procedure was later developed into what the authors termed "teach-aloud" and has implications for teaching in education (Mathews et al, 1989).

Although the experiment reported here does not encourage participants to explicitly explain their understanding to someone else, it could be argued that the two stages of the task in the silent condition encourage participants to provide an explanation of their understanding. Whilst thinking aloud participants may simply be trying to understand the material presented to them for themselves but when asked in a separate stage to tell the experimenter what their understanding of the graphs is they need to then ensure the information provided can be understood by someone else, which is similar to the written condition (Klein, 1999).

However, this effect can be balanced by including the second stage of the silent condition in the think aloud condition. Therefore, in order to test whether it was communicating understanding that resulted in a performance improvement in the silent condition, a communication condition was included as an experiment. In this condition participants were required to interpret the graphs whilst thinking aloud, consistent with the earlier experiments. Once they felt they had understood the graphs as much as possible they verbalized their interpretation to the experimenter

Method

Participants

Fifteen undergraduate Psychology students (13 female, 2 male) from the University of Huddersfield volunteered to take part in the experiment for which they were paid £5 in vouchers. The age of participants ranged from 19.1 to 29.7 years with a mean of 19.9 years (SD = 1.8). All participants were in the first year of a three year Psychology degree.

Materials, Design and Procedure

The experiment was carried out using the same equipment and the same procedure as Experiment 6, (silent condition). Participants were instructed the experiment consisted of two stages – in the first "think aloud" stage they were to think aloud whilst interpreting the graph they were viewing. In the second "talking" stage they were to tell the experimenter what they had understood about the graph.

Results

Results were similar to other think aloud conditions – the majority of the sample was classified pre-elementary. The high rate of pre-elementary performance found in the think aloud condition in Experiment 5 was replicated with 53% of the sample classified as pre-elementary in this condition. Similar to the think aloud condition in Experiment 5 and earlier experiments, the majority of participants were classified as pre-elementary, and a smaller proportion elementary and intermediate. Figure 23 displays the number of users in each comprehension category, alongside the results of the think aloud condition in Experiment 5 and 6 for comparison.

Including a stage where participants were required to communicate their understanding resulted in some benefits. Surprisingly, despite a similar rate of pre-elementary performance to that of the think aloud condition, 33% of participants interpreted all six trials correctly in this line graph condition. This compares favourably to 13% of line graph readers in the think aloud condition. However, this association was not significant (Fisher's Exact test, p = .39).

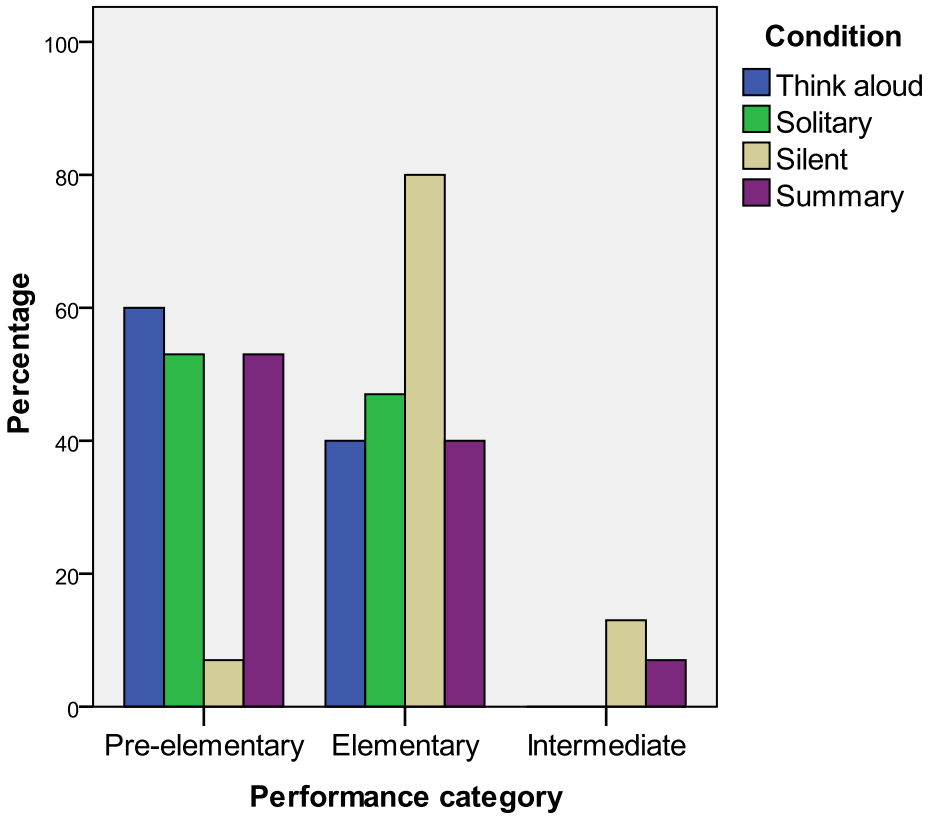


Figure23. Percentage of line graph users in the three performance categories - Experiments 5-7

A chi-square test revealed that the association between the number of participants classified as pre-elementary in the communication and think aloud line graph condition was not significant (chi-square = .14 df = 1; p = .71).

A comparison of the number of correct trials between the think aloud, solitary, silent and summary condition also revealed that the silent condition resulted in a significant increase (Kruskal-Wallis H = 7.10, df = 3, p < .05) in the number of correctly interpreted trials (mean rank = 40.03) compared to the think aloud condition (mean rank = 23.83), solitary condition (mean rank = 26.20), and summary condition (mean rank = 31.93).

Four post-hoc Mann Whitney U tests (with alpha levels Bonferroni adjusted to .0125) revealed the significant difference to be between the silent and think-aloud line graph conditions ($p < 0.01$), but not between the solitary condition ($p = .03$) nor between the summary condition ($p = .33$).

Discussion

The results of this experiment reveal that it was not the second stage of the silent condition which resulted in a drop in pre-elementary performance. Communicating their understanding to another person did not result in improvements in participants' performance when comparing number of correct trials between this condition and the think aloud condition. Therefore, the findings of this study allow me to conclude that it is the demands of verbalization which result in a detriment in performance in the line graph condition.

Although research has demonstrated that communicating understanding improves performance on various measures, this research has utilized specific types of instruction known to result in performance improvements. For example, although it could be argued requiring participants to communicate understanding is similar to eliciting self-explanations like Chi, Leeuw, Chiu and Lavancher (1994) did, this experiment was not designed for this purpose. Participants were encouraged to explicitly communicate their interpretation to someone else rather than explain their understanding which can account for why no performance improvement was observed. Similarly, in the study conducted by Stanley et al (1989) participants knew they were tutoring other students with the aim of guiding them to the correct response which makes task demands different to the ones present in this experiment.

Chapter 9

General discussion

Introduction

The primary aim of this research was to investigate limitations in students' comprehension of statistical graphs and identify ways in which to overcome these limitations. A series of studies was conducted to investigate the various factors influencing graph comprehension with the aim of developing theory to inform graph design. The focus of the experiments was to investigate two of the three components identified in the graph comprehension literature as predicting how well a graph will be understood in an applied setting. Specifically, the effects of graph format and the nature of the interaction with the diagram were investigated to determine how they affected novice users' interpretation of three-variable bar and line graphs.

In this chapter I outline key findings from this research and the implications they have for graph design and the nature of the interaction users' benefit from. First, I consider the results of each experiment in turn and provide methodological evaluation of how the research was conducted. As this research is applied to educational learning the implications of the findings are discussed for practice and education. Then I discuss whether the verbal protocol method is an appropriate methodology to employ when assessing conceptual understanding of material in educational settings. Finally I use the results of this research to make recommendations for graph design and consider how the research finding could be developed in further research.

Summary of research and key findings

The first aim of the research was to identify how informationally equivalent three-variable interaction graphs are understood by students who, as part of their studies, are required to interact with them. An analysis of informationally equivalent bar and line graphs revealed significantly poorer performance for line graph users than for users of bar graphs which error analysis determined was due to an imbalance in how the

independent variables were represented by the plotted lines. Further analysis suggested that the underlying cause of this imbalance was the action of Gestalt principles of perceptual organization which drew users' attention to specific graphical elements and influenced the subsequent identification of variables.

This finding motivated the second goal of the project – to test the hypothesis that Gestalt principles were the underlying cause of the observed poor performance and to determine whether these principles could be utilized to design a more balanced line graph that represented the independent variables more equally. In Experiment 2, this balance was attempted by adding the features of bar graphs to the line graphs by including 'drop-down' lines to explicitly connect the data points to the x axis values. Although this modified design did result in a modest improvement in performance (thereby providing some support for the hypothesis), analysis of verbal protocols showed that participants found the novel design too visually complex and confusing. This led to a reappraisal of the design and an attempt to produce a new, less cluttered graph which utilized the Gestalt principles already in use in the original line graph.

Representational balance was achieved by employing the Gestalt principle of similarity to allow users to associate the plot points with values of both independent variables. The resulting graphs were tested in Experiment 3, the results of which clearly demonstrated a significant improvement in comprehension performance, to the same level as the bar graphs in Experiment 1. These three experiments provide strong support for the claim that graph design is a key factor which determines readers' conceptual understanding of the relationships depicted and that modifying the design of graphs can result in significant improvements in readers' ability to interpret them.

Having established that design can influence interpretation, attention was turned to the second major determinant of comprehension performance – interaction type and the requirements of the task. In the first three experiments, participants were required to attempt to interpret the graph until they were satisfied they had understood it while concurrently thinking aloud to the experimenter who was present. Previous studies have shown however that different forms of interaction with represented material may increase the depth of processing of the material which results in improved comprehension (e.g., Benton, Kiewra, Whitfill and Dennison, 1993). This prompted the second goal of the project – to determine whether performance with the original lines can be raised to the level of bar graphs simply by changing the mode of interaction.

Experiment five investigated the effect of writing on the nature and quality of interpretations for bar and line graphs. In line with previous studies the experiment showed that writing an interpretation resulted in a significant improvement in the comprehension of line graphs compared to the original mode of interaction. This improvement suggests that verbal and written interactions provide different indications of graph comprehension.

The final three experiments were designed to identify the possible cause of the differences observed in Experiment 5. A review of the literature suggested two possible candidate causes for the reduced performance in the original experiment: (a) the inhibitory effect of the presence of the experimenter, and (b) the increased cognitive demands imposed by thinking aloud (so-called 'reactivity effects'). These hypotheses were tested in Experiments 6 and 7 respectively. The results of these experiments did not support the experimenter presence hypothesis but did support the suggestion that reactivity effects were the underlying cause of the reduced performance in the original line graph condition.

Methodological evaluation

The main method of data collection used throughout to investigate the research question was the verbal protocol method. The reason for this was because this particular methodology allowed me to trace the cognitive processes leading to the errors students make whilst interpreting graphs. This in depth analysis of cognitive processes underlying graph comprehension allowed for the creation of a novel graph design which was successful in reducing pre-elementary performance. Although a change in methodology (eliciting written responses instead of a protocol) revealed the verbal protocol method was not necessarily appropriate for assessing conceptual understanding of these types of graphs, the findings suggest the benefits of employing this method outweighed the potential drawbacks.

One possible criticism of how the experiments were conducted to answer the research question is that comparing findings across experiments is not appropriate and lacks experimental rigor. However, the way in which the studies were conducted counteracts this criticism. Firstly, the high rate of pre-elementary performance in the line graph condition has been demonstrated consistently in a number of experiments.

Peebles and Ali (2009) found this effect and Experiment 1 confirmed the effect was a lot more pronounced in an undergraduate student population. In addition to this, Experiment 4 demonstrated an almost identical pattern of results when a third year student sample was used. Experiment 5 established this effect remains when using stimuli depicting different content. Therefore, although results were compared across experiments, the high rate of pre-elementary performance in the line graph condition has unequivocally been demonstrated.

In addition to this the sample used in the experiments was drawn from the same population for those conditions which were compared. First and second year psychology students were used in experiments one, two and three and statistical analysis revealed there were no significant variations between foundation and intermediate level students. Experiment 4 only used a sample of third year students and conditions were not compared. Finally, Experiment five, six and seven only used first year students. Furthermore, to ensure analysis and scoring of transcripts was rigorous 25% of transcripts from all the conditions in experiments one, two, three and five were scored by an independent researcher and inter-rater reliability was high (above 85% in all cases).

Another possible confound present is the time difference between conditions to complete the task in experiment five. Those students who provided a written response took longer to complete the task than those who provided a verbal protocol. This can partially be accounted for by the time it takes to write a response, but it could be argued that the additional time resulted in additional processing which can account for the improvement in performance. However, as this is an open ended task restricting the time to provide a response would have resulted in a greater confound, and both conditions were identical in instruction – participants were told to take as long as they need to provide a response which reflected their understanding of the graphs.

Furthermore, analysis of verbal protocols indicated additional time spent on the task did not result in any improvements. Those participants who took longer to provide a protocol simply repeated incorrect assumptions already stated and the additional time seemed to strengthen their erroneous beliefs about the relationships the graphs were depicting. Further evidence that it was not time on task which resulted in the observed performance improvements comes from the silent condition. Time spent on task was not as long as

the written condition and the same performance improvements were observed as the written line graph condition. Therefore, findings from Experiment 7a are consistent with the claim that time spent on task does not improve performance from increased depth of processing of material.

Implications for practice and education

The results of this research have important implications for line graph use when a particular audience is required to interact with them. Research has consistently demonstrated students will struggle to comprehend relationships accurately when attempting to interpret graphs, despite this skill being a key requirement of the course (Bowen, Roth & McGinn, 1999, Friel, Curcio and Bright, 2001, Carlson et al, 2002). In an attempt to overcome these difficulties guidelines for effective construction of graphical displays based on gestalt principles of perceptual organisation have been proposed to ease interpretation (Kosslyn, 1989, Pinker, 1990, Shah, Mayer, and Hegarty, 1999). However, there has been a limited amount of research investigating students' conceptual understanding of these graph types and how gestalt principles operate in these displays to shape viewers' understanding of the graph. This applied research contributes to our understanding of how to effectively incorporate gestalt principles of perceptual organisation in graphs to ease interpretation.

This is particularly important in educational settings where novice readers are required to interact with such data with little or no instruction. The colour match design proposed here eases interpretation by facilitating association of pattern to referents. The effect of this is that pre-elementary performance is reduced to that of the bar graph users. However, based on the findings of this research graph instruction needs to play a key role in educational settings as graph design will only increase performance to elementary – and in a few cases – intermediate level. In order for students to become advanced users explicit instruction is necessary to enrich schemas so that they can identify patterns and the relationships they signify with ease (Pinker, 1990).

Implications for use of the verbal protocol method

The results of this series of experiments have important implications for the use of the verbal protocol method when attempting to assess comprehension of material. In the three experiments using verbal protocols, the demands of verbalization interfered with comprehension processes, increasing the detrimental effect of Gestalt principles of perceptual organization in line graphs. This supports the findings of Russo, Johnson and Stephens (1989) and Schooler, Ohlsson and Brooks (1993), who also demonstrated similar detrimental effects.

There are a number of responses to these findings. It could be argued that they call into question the use of the think aloud method to understand cognitive processes as one may never be sure that reactivity effects are occurring. This response may be too extreme however. An alternative response is to ensure that reactivity effects are not occurring by including of a silent control condition in the initial stages of a research project (as recommended by Russo, Johnson & Stephens, 1989).

Although a silent control condition may validate the use of the think aloud method, it does not necessarily ensure that this method is the most appropriate for assessing conceptual understanding of material. It may be the case that other methods (e.g., producing written accounts) are more accurate indicators of users' knowledge and abilities by providing better opportunities for users to interact with the diagram, either through different problem solving strategies or goals or through similar goals without the additional demands of verbalization. This is a particularly important issue in educational research contexts in which novice users' conceptual understanding of material is assessed in an attempt to identify ways in which to improve their understanding.

It would appear that the theory of protocol generation proposed by Ericsson and Simon (1993) is unable to predict, in some cases at least, when the verbal protocol method can result in reactivity effects. In addition, our understanding of the verbal protocol method remains limited due to the small number of studies reporting reactivity effects. Based on the results of this research and those of Russo, Johnson & Stephens (1989) and Schooler, Ohlsson and Brooks (1993), the conclusion to be drawn is that empirical checks for

reactivity effects are necessary in order to eliminate the possibility that they arise and affect the behavior being measured.

Recommendations for graph and interaction design

The final consideration relates to recommendations concerning which design is appropriate to employ when communicating the results of factorial designs. As there were three conditions in this experiment which resulted in a performance improvement (colour match graph, written interpretation, and silent condition) it is important to identify which in general would be the most appropriate to employ for these types of tasks. Although Experiment 7 revealed that the high rate of pre-elementary performance observed in the line graph condition was due to the act of verbalizing, one of the performance improvements observed emerged due to the novel colour match design tested using the verbal protocol method.

The research literature is broadly consistent on the principles to consider when making recommendations about which type of display to employ in different contexts. Wherever possible the number of inferential processes should be minimized and the number of pattern matching processes maximized (Parkin, 1983, cited by Pinker, 1990, Kosslyn, 1989, 1994, Shah and Carpenter, 1995, Shah Meyer and Hegarty, 1999). The earlier experiments revealed that the standard line graph display employed in the literature does not follow this recommendation; pattern matching processes are only available for the variable plotted in the legend.

However, the colour match graph addresses these issues by ensuring that pattern matching processes are available to match the pattern at the centre of the display to both levels of each independent variable. Consistent with the explanations proposed here for why the colour match graph was effective in reducing erroneous interpretations, cognitive load theory (Sweller, 1994) and information display guides (Zhang, 1996) assume that information presented to learners should be structured to eliminate any avoidable load on working memory. Similar to assumptions proposed by the proximity compatibility principle (Carswell and Wickens, 1995) research derived from cognitive load theory and display design makes recommendations for efficient display design to encourage rapid and effortless processing. For example, Sweller (1994)

recommends instructional material should not require readers to split their attention between diagrams and text. Instead, text should be ‘physically close’ in space to the diagram so search and locating operations are reduced.

Although the written condition resulted in a considerable improvement in performance, this task is perhaps not appropriate to recommend from a speed/accuracy trade-off perspective. This is because although the number of correct responses was high, this condition also resulted in the lengthiest response times. The minimum amount of time students took to complete the task was 25 minutes, the longest 40 minutes. Even when the time to write sentences is factored in, this is a considerable amount of investment required to interpret graphs in this condition, which have been estimated to take 30 seconds (Shah, 2002).

Although accuracy increased dramatically, it is unrealistic to expect students to spend a considerable amount of time writing out an interpretation of graphs they see in textbooks or research literature. Furthermore, this option may not always be available. When graphs are presented during talks (e.g., lectures, conferences) the speed with which information is presented and the pace of talking does not allow the audience to deliberate over information for long periods of time.

The other option is the silent condition, where performance was on par with the written condition and elementary-level performance was relatively high. Participants in this condition took slightly longer to complete the task than the think aloud condition but considerably less time than those in the written condition. Therefore, findings from this research suggest the best possible recommendations would be to employ the colour match graph and suggest novice users simply read the graphs to themselves silently. Employing the colour match graph would increase the number of pattern match processes, thus easing interpretation for users who are still not familiar with graphical conventions. This design-based solution provides the appropriate representational features to support correct associations between pattern and referents which promotes accurate interpretation and the development of pattern recognition schemas.

Further research

A number of findings have emerged from the experiments conducted to answer the research question which could be explored in further research. Firstly, the colour match graph resulted in a significant performance improvement, demonstrating deign of visual displays plays a crucial role in basic processes involved in comprehension of this type of material. Further research could explore the strength of this effect by employing different methodological techniques to assess comprehension (for e.g., question answer tasks, written responses). In addition to this the gestalt principles employed in this line graph display to increase pattern matching processes could be applied to other types of displays to test whether this results in improvement of novices understanding of such displays.

However, the facilitatory effects of this novel graph design may be limited to specific types of graphs which only depict a certain number of variables. Literature investigating information processing capacity limitations (Halford, Baker, McCredden and Bain, 2005) investigated how many variables participants could process together when interpreting graphically displayed statistical interactions. They found accuracy and response time decreased significantly from three-way to four-way interactions. Performance on a five-way interaction was at chance level. These findings demonstrate how when processing capacity is exceeded the ability to interpret relations depicted in graphs is compromised. Therefore, a novel design such as the colour match graph may only result in an improvement for graphs depicting up to three-way and perhaps four-way interactions.

Secondly, the written condition revealed performance improvement measures. A review of the writing to learn literature reveals inconsistent findings concerning whether the act of writing improves conceptual understanding of material (Klein, 1999). These findings suggest further research is necessary to identify when writing can result in performance improvements. Such research would provide valuable guidance to educators to which is the appropriate type of interaction with material students should engage in.

Finally, another major finding which is scarce in the literature is the reactivity effects that emerged from employing the verbal protocol method to investigate the research question. These findings indicate further

research is required to determine when the verbal protocol method can alter cognitive processes involved in task completion. The results of experiment seven further strengthen Russo, Johnson and Stephens' (1989) conclusion that task demands interact with the demands of verbalization to influence cognitive processes involved in undergoing a task. Therefore, recommendations for further research would be to identify tasks in which reactivity effects can emerge from employing the verbal protocol method so potential bias in findings can be avoided.

References

Ackerman, J. M. (1993). The promise of writing to learn. *Written Communication*, 10, 334–370.

Ahlum-Heath, M. E., & DiVesta, F. J. (1986). The effect of conscious controlled verbalization of cognitive strategy on transfer in problem solving. *Memory and Cognition*, *14*, 281–285.

Anzai, Y. (1991). Learning and use of representations for physics expertise. In K. A. Ericsson & Smith, Eds. *Towards a General Theory of Expertise.* pp 64 – 92. Cambridge: Cambridge University Press.

Applebee, A. N. (1984). Writing and reasoning. *Review of Educational Research*, 54(4), 577–596.

Aron, A., Aron, E. N., & Coups, E. J. (2006). *Statistics for psychology* (4th ed.). London: Pearson.

Bell, A., Brekke, G., & Swan, M. (1987). Misconceptions, conflict and discussion in the teaching of graphical interpretation. In J. D. Novak (Ed.), *Proceedings of the second international seminar: Misconceptions and educational strategies in science and mathematics* (Vol. 1, pp. 46–58). Ithaca, NY: Cornell University.

Benton, S. L., Kiewra, K. A., Whitfill, J. M., and Dennison, R. (1993). Encoding and external storage effects on writing processes. *Journal of Educational Psychology* 85: 267-280.

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Berg, C., & Phillips, D. G. (1994). An investigation of the relationship between logical thinking structures and the ability to construct and interpret line graphs. *Journal of Research in Science Teaching*, *31*, 323–344.

Berry, D. C, & Broadbent, D. E. (1984). On the relationship between task performance and associated verbalizable knowledge. *Quarterly Journal of Experimental Psychology, 36A,* 209-231

Bertin, J. (1983). *Semiology of Graphics*. University of Wisconsin Press, London.

Bond, C. F. (1982). Social facilitation: A self-presentational view. *Journal of Personality and Social Psychology, 42,* 1042-1050

Bond, C. F., & Titus, L. J. (1983). Social facilitation: A meta-analysis of 241 studies. *Psychological Bulletin*, 94, 265-292

Boyd, D., & Bee, H. (2006). *Lifespan development* (4th ed.). Boston: Allyn and Bacon.

Britton, J. (1978). The composing processes and the functions of writing. In C. R. Cooper & L. Odell (Eds.), *Research on composing: Points of departure* (pp. 13–28). Urbana, IL: NCTE.

Brinton, W. C. (1914). *Graphic methods for presenting facts*. New York: Engineering Magazine.

Carlson, M., Jacobs, S., Coe, E., Larsen, S., & Hsu, E. (2002). Applying covariational reasoning while modeling dynamic events: A framework and a study. *Journal for Research in Mathematics Education*, 33(5), 352–367.

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

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

Carswell, C. M., & Wickens, C. D. (1987). Information integration and the object display: An interaction of task demands and display superiority. *Ergonomics*, 30 (3), 511-527.

Carswell, C. M., & Wickens, C. D. (1990). The perceptual interaction of graphical attributes: Configurality, stimulus homogeneity, and object interaction. *Perception & Psychophysics*, *47*, 157–168.

Carswell, C. M., & Wickens, C. D. (1996). Mixing and matching lower-level codes for object displays: Evidence for two sources of proximity compatibility. *Human Factors*, *38* (1), 1–22.

Chi, M. T. H., de Leeuw, N., Chiu, M., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, *18*, 439-477

Cleveland, W. S. (1985). *The elements of graphing data*. Monterey, CA: Wadsworth.

Cleveland, W. S., & McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association, 79*, 531–554.

Cottrell, N. B., Wack, D. L., Sekerak, G. J., & Rittle, R. H. (1968). Social facilitation of dominant responses by the presence of an audience and the mere presence of others. *Journal of Personality and Social Psychology, 9*, 245-250.

Croxton, F.E. (1927), "Further Studies in the Graphic Use of Circles and Bars II: Some Additional Data," *Journal of the American Statistical Association*, 22, 36-39

Croxton, F, E., and Stein, H (1932), "Graphic Comparisons by Bars, Squares, Circles and Cubes," *Journal of the American Statistical Association*, 27, 54-60.

Croxton and Stryker, R.E. (1927), "Bar Charts Versus Circle Diagrams," *Journal of the American Statistical Association*, 22, 473-482.

Crutcher, R. J. (1994). Telling what we know: The use of verbal report methodologies in psychological research. *Psychological Science*, 5, 241–244.

Culbertson, H. M., & Powers, R. D. (1959). A study of graph comprehension difficulties. *Audio-Visual Communication Review, 7*, 97–110.

Cumming, A. (1989). Writing expertise and second- language proficiency. *Language Learning*, 39, 81- 141.

Curcio, F. R. (1987). Comprehension of mathematical relationships expressed in graphs. *Journal for Research in Mathematics Education, 18*, 382–393.

Dillashaw, F. G., & Okey, J. R. (1980). Test of the integrated science process skills for secondary science students. *Science Education, 64*, 601–608.

Dancey, C. P., & Reidy, J. (2008). *Statistics without maths for psychology* (4th ed.). Essex: Pearson.

Eells, W. C. (1926). The relative merits of circles and bars for representing component parts. *Journal of the American Statistical Association, 21*, 119–132.

Emig, J. (1977). Writing as a mode of learning. *College Composition and Communication*, 28, 122–128.

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data.* Cambridge, MA: MIT Press.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Revised ed.). Cambridge, Mass: MIT Press.

Eysenck, M. W., & Keane, M. T. (2005). *Cognitive psychology: A student's handbook* (5th ed.). Hove: Psychology Press.

Few, S. (2010). *BP tries to mislead you with graphs.* Available: http://flowingdata.com/2010/05/26/bp-tries-to-mislead-you-with-graphs/. Last accessed 15th Feb

Field, A. (2009). *Discovering statistics using SPSS* (4th ed.). London: Sage.

Flower, L., & Hayes, J. R. (1980). The cognition of discovery: Defining a rhetorical problem. *College Composition and Communication*, 31, 21–32.

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication*, 32, 365–387.

Freedman, E.G., and Smith, L.D. (1996).The role of data and theory in covariation assessment: Implications for the theory-ladenness of observation. *Journal of Mind and Behaviour*. 17: 321–343.

Friel, S. N., Curcio, F. R., & Bright, G. W. (2001). Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in Mathematics Education*, *32*, 124–158.

Gal, I. (2002). Adult statistical literacy: Meanings, components, responsibilities. *International Statistical Review* 70(1): 1-25.

Galbraith, D. (1992). Conditions for discovery through writing. *Instructional Science*, 21, 45–72.

Geen, R. G. (1983). Evaluation apprehension and the social facilitation/inhibition of learning. *Motivation and Emotion*, *7*(2), 203-212.

Gray, W. D., & Altmann, E. M. (2001). Cognitive modeling and human-computer interaction. In W. Karwowski (Ed.), *International encyclopedia of ergonomics and human factors* (pp. 387-391). New York: Taylor & Francis Ltd.

Greene, S. (1993). The role of task in the development of academic thinking through a reading and writing in a college history course. *Research in the Teaching of English* 27: 46-75.

Greene s., & Ackerman, J. M. (1995). Expanding the constructivist metaphor: A rhetorical perspective on literary practice. *Review of Educational Research,* 65(4), 383–420.

Guerin, B. (1986). Mere presence effects in humans. *Journal of Experimental Social Psychology*, 22, 38-77.

Halford, G. S. Baker, R. McCredden, J. E and Bain, J. D. (2005) "How many variables can humans process?" *Psychological Science*, vol. 16, no. 1, pp. 70–76.
Harris, R. L. (1999). *Information graphics: A comprehensive illustrated reference*. Atlanta, USA: Management Graphics.

Hayes, J. R., & Flower, L. (1983). Uncovering cognitive processes in writing. In P. Mosenthal, L. Tamor, & S. Walmsley (Eds.), *Research in writing: Principles and methods* (pp. 207-220). New York: Longman

Hoffman, R. R. (1991). Human factors psychology in the support of forecasting: The design of advanced meteorological workstations. W*eather and Forecasting*, 6, 98-110.

Howitt, D., & Cramer, D. (1998*). Introduction to statistics in psychology* (4th ed.). Essex: Pearson.

Huguet, P., Galvaing, M., Monteil, J., & Dumas, F. (1999). Social presence effects in the Stroop task: Further evidence for an attentional view of social facilitation. *Journal of Personality and Social Psychology, 77,* 1011–1025.

Kahneman, D. (1973). *Attention and effort.* Englewood Cliffs, N.J.: Prentice-Hall.

Klein, P. D. (1999). Reopening inquiry into cognitive processes in writing-to-learn. *Educational Psychology Review*, 11, 203-270.

Koedinger, K. R., & Anderson, J. R. (1990). Abstract planning and perceptual chunks: Elements of expertise in geometry. *Cognitive Science*, 14, 511–550.

Kosslyn, S. M. (1989). Understanding charts and graphs. *Applied Cognitive Psychology*, 3, 185-226.

Kosslyn, S. M. (1994). *Elements of graph design*. New York: W. H. Freeman.

Kosslyn, S. M. (2006). *Graph design for the eye and mind*. New York: Oxford University Press.

Kosslyn, S. M. & Chabris, C. (1993, September/October). The mind is not a camera; the brain is not a VCR. *Aldus Magazine*, 33-36

Langdridge, D., & Hagger-Johnson, G. (2009). *Introduction to research methods and data analysis in psychology*. Harlow: Pearson Prentice Hall.

Langer, J. A. (1986). *Children reading and writing: Structures and strategies*. Norwood, NJ: Ablex.

Larkin, J. H. (1989). Display-based problem solving. In D . Klahr & K . Kotovsky Y , Eds . *Complex Information Processing: The Impact of Herbert A . Simon.* pp . 319 – 341. Hillsdale, NJ: Lawrence Erlbaum Associates.

Larkin, J. H., McDermott, J., Simon, D. P., & Simon, H. A. (1980). Expert and novice performance in solving physics problems. *Science*, 208, 1335–1342.

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

Leinhardt, G., Zaslavsky, O., & Stein, M. K. (1990). Functions, graphs, and graphing: Tasks, learning, and teaching. *Review of Educational Research, 60,* 1–64.

Lewandowsky, S., & Behrens, J. T. (1999). Statistical graphs and maps. In F. T. Durso, R. S. Nickerson, R. W. Schvaneveldt, S. T. Dumais, D. S. Lindsay, & M. T. H. Chi (Eds.), *Handbook of applied cognition* (pp. 513- 549). Chichester, UK: Wiley.

Lowe, R. (1993). *Successful instructional diagrams*. London: Kogan Page.

Lowe, R. (2000). *Visual literacy and learning in science*. Eric Digest, Report EDO-SE-00-02.

Maichle, U. (1994). Cognitive processes in understanding line graphs. In W. Schnotz & R. W. Kulhavy (Eds.), *Comprehension of graphics* (pp. 207–226). New York: Elsevier Science.

Markus, H. (1978). The effect of mere presence on social facilitation: An unobtrusive test. Journal *of Experimental Social Psychology*, 14, 389-397.

Mathews, R. C., Buss, R. R., Stanley, W. B., Blanchard-Fields, F., Cho, J. R., & Druhan, B. (1989). Role of implicit and explicit processes in learning from examples: A synergistic effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 1083-1100

McCrindle, A. R., and Christensen, C. A. (1995). The impact of learning journals on metacognitive and cognitive processes and learning performance. *Learning and Instruction 5:* 167-185.

Mooney, E. S. (2002). A framework for characterizing middle school students' statistical thinking. *Mathematical Thinking and Learning*, *4*(1), 23-63.

Murnane, Richard; John Willett, and Frank Levy. (1995). The Growing Importance of Cognitive Skills in Wage Determination, *Review of Economics & Statistics*. 77:2, pp. 251-66**.**

No author. (2009). *Surface pressure forecast.* Available: http://www.metoffice.gov.uk/weather/uk/surface_pressure.html. Last accessed 23 September 2009

No author. (2009). *Improving data visualisation for the public sector.* Available: http://www.improving-visualisation.org/vis/id=49. Last accessed 15th Feb 2011

Nathan, M.J. Mertz, K., & Ryan, B. (1993). *Learning through self-explanation of mathematical examples: Effects of cognitive load.* Paper presented at the 1994 Annual Meeting of the American Educational Research Association.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice Hall.

Newell, G. E. (1984). Learning from writing in two content areas: A case study/protocol analysis. *Research in the Teaching of English*, 18, 265–287.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84,* 231-259.

Olson, D. R. (1977). From utterance to text: The bias of language in speech and writing. *Harvard Educational Review*, 47, 257–281.

Ong, W. J. (1982). *Orality and literacy*. New York: Methuen Inc.

Padilla, M. J., McKenzie, D. L., & Shaw, E. L. (1986). An examination of the line graphing ability of students in grades seven through twelve. *School Science and Mathematics, 86,* 20–26.

Palmer, S. E., & Rock, I. (1994). Rethinking perceptual organization: The role of uniform connectedness. *Psychonomic Bulletin & Review, 1*, 29–55.

Parkin, L. (1983). *A comparison of various graph labelling methods in the context of Gestalt organizing principles*. Unpublished manuscript. Consulting Statisticians, Inc., Wellesley, MA.

Paulos, J. (1988). *Innumeracy: Mathematical illiteracy and its consequences*. New York: Hill and Wang.
Payne J W (1994) Thinking Aloud: Insights into Information Processing. *Psychological Science*, vol 5, no 5, pp 241-248

Peebles, D. (2008). The effect of emergent features on judgments of quantity in configural and seperable displays. *Journal of Experimental Psychology: Applied, 14*(2), 85– 100.

Peebles, D., & Ali, N. (2009). Differences in comprehensibility between three-variable bar and line graphs. In Proceedings of the thirty-first annual conference of the cognitive science society (pp. 2938 - 2943). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Penrose, A. M. (1992). To write or not to write: Effects of task and task interpretation on learning through writing. *Written Communication*, 9, 465–500.

Pereira-Mendoza, L., & Dunkels, A. (1989). Stem-and-leaf plots in the primary grades. *Teaching Statistics, 13*, 34–37.

Phillips, R. (1997). Can juniors read graphs? A review and analysis of some computer-based activities. *Journal of Information Technology for Teacher Education* 6: 49-58.

Pinker, S. (1990). A theory of graph comprehension. In R. Freedle (Ed.), *Artificial intelligence and the future of testing* (pp. 73–126). Hillsdale, NJ: Lawrence Erlbaum Associates.

Pirolli, P.L. & Recker, M. (1994). Learning strategies and transfer in the domain of programming. *Cognition and Instruction, 12*, 235-275.

Preece, J., & Janvier, C. (1992). A study of the interpretation of trends in multiple curve graphs of ecological situations. *School Science and Mathematics*, 92, 299-306.

Pugalee, D. (2004). A comparison of verbal and written descriptions of students' problem solving processes. *Educational Studies in Mathematics*, 55, 27–47.

Ratwani, R.M. & Trafton, J.G. (2008). Shedding light on the graph schema: Perceptual features vs. invariant structure. *Psychonomic Bulletin & Review*. 15(4), 757-762.

Ratwani, R. M., Trafton, J. G., & Boehm-Davis, D. A. (2008). Thinking graphically: Connecting vision and cognition during graph comprehension. *Journal of Experimental Psychology: Applied*, 14(1), 36–49.

Reece M. & Whitman. R (1962) Expressive movements, warmth, and verbal reinforcement. *Journal of Abnormal and Social Psychology* Volume 64, Issue 3, 234-236.

Rivard, L. P. (1994). A review of writing-to-learn in science: Implications for practice and research. *Journal of Research in Science Teaching* 31: 969-983.

Rivera-Batiz, Francisco L. (1992) Quantitative Literacy and the Likelihood of Employment among Young Adults. *Journal of Human Resources* 27, 313-328.

Rosenthal, R. (1976) *Experimenter effects in behavioural research*. New York: Irvington.

Roth, W. -M., & McGinn, M. (1997). Graphing: Cognitive ability or practice? *Science Education, 81*, 91–106.

Russo, J. E., Johnson, E. J., & Stephens, D. L. (1989). The validity of verbal protocols. *Memory and Cognition*, 17, 759–769.

Schmid, C. F., & Schmid, S. E. (1979). *Handbook of graphic presentation* (2nd ed.). New York: Wiley.

Schooler, J. W., Ohlsson, S., & Brooks, K. (1993). Thoughts beyond words: When language overshadows insight. *Journal of Experimental Psychology*: General, 122, 166–183.

Schumacher, G. M., and Nash, J. G. (1991). Conceptualizing and measuring knowledge change due to writing. *Research in the Teaching of English* 25: 67-96.

Schutz, H. G. (1961). An evaluation of formats for graphic trend displays—experiment II. *Human Factors*, *3*, 99–107.

Sensenbaugh, R. (1989). Writing across the curriculum. *Journal of Reading* 32: 462-465.

Shah, P. (1995). *Cognitive Processes in Graph Comprehension*, Unpublished doctoral dissertation.

Shah, P., & Carpenter, P. A. (1995). Conceptual limitations in comprehending line graphs. *Journal of Experimental Psychology: General*, *124*, 43–62.

Shah, P., & Freedman, E. (2009). Bar and line graph comprehension: An interaction of top-down and bottom-up processes. *Topics in Cognitive Science*, doi: 10.1111/j.1756-8765.2009.01066.x.

Shah, P., and Hoeffner, J. (2002), "Review of Graph Comprehension Research: Implications for Instruction," *Educational Psychology Review*, 14, 47–69.

Shah, P., & Miyake, A. (2005). *The Cambridge handbook of visuospatial thinking*. New York: Cambridge University Press

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

Shaughnessy, J. M., Garfield, J., & Greer, B. (1996). Data handling. In A. J. Bishop, K. Clements, C. Keitel, J. Kilpatrick, & C. Laborde (Eds.), *International handbook of mathematics education* (pp. 205-237). Boston, MA: Kluwer**.**

Simcox, W. A. (1983). *A perceptual analysis of graphic information processing*. Unpublished doctoral dissertation, Tufts University, Medford, MA.

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

Smith, L. D., Best, L. A., Stubbs, D. A., Johnston, J., & Archibald, A. B. (2000). Scientific graphs and the hierarchy of the sciences: A Latourian survey of inscription practices. *Social Studies of Science, 30*, 73–94.

Stanley, W. B., Mathews, R. C., Buss, R. R., & Kotler-Cope, S. (1989). Insight without awareness: On the interaction of verbalization, instruction and practice in a simulated process control task. *The Quarterly Journal of Experimental Psychology, 41A*, 553-577.

Stevenson. H., W & Allen, S (1964). Adult performance as a function of sex of experimenter and sex of subject. *Journal of Abnormal and Social Psychology*. 68 (2), 214-216

Sweller, J. (1994). Cognitive load theory, learning difficulty and instructional design. *Learning and Instruction*, 4, 295-312

Tabachneck-Schijf, H. J. M., Leonardo, A.M. & Simon, H. A. (1997). CaMeRa: a computational model of multiple representations. *Cognitive Science*, 21, 305-350.

The Investor. (2008). *The Barclays share price and the credit crunch.* Available: http://monevator.com/2008/11/13/the-barclays-share-price-and-the-credit-crunch/. Last accessed 16th Feb 2011.

Trafton, J. G. Kirschenbaum, S. S. Tsui, T. L. Miyamoto, R. T Ballas, J. A. and. Raymond, P. D. (2000) "Turning Pictures into Numbers: Extracting and Generating Information from Complex Visualizations," *International Journal of Human-Computer Studies*, vol. 53, pp. 827-850.

Treisman, A. M. (1985). Pre-attentive processing in vision. *Computer Vision, Graphics, and Image Processing*, 31, 156 - 177

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

Tversky, B. (2001). Spatial schemas in depictions. In M. Gattis (Ed.), *Spatial schemas and abstract thought*. Cambridge MA: MIT Press.

Tversky, B. , Kugelmass, S. & Winter, A. (1991). Cross-cultural and developmental trends in graphic productions. *Cognitive Psychology*, 23, 515-557

Tynjala, P. (1999). Towards expert knowledge? A comparison between a constructivist and a traditional learning environment in university. *International Journal of Educational Research,* 31(5), 357–444.

Vernon, M. D. (1946). Learning from graphical material. *British Journal of Psychology, 36,* 145–158.

Von Huhn, R. (1927), "Further Studies in the Graphic Use of Circles and Bars I: A Discussion of Eells' Experiment," *Journal of the American Statistical Association*, 22, 31-36.

Wallman, K.K. (1993). Enhancing statistical literacy: Enriching our society. *Journal of the American Statistical Association,* 88, 1-8.

Watson, J. and Callingham, R. (2003). Statistical literacy: A complex hierarchical construct. *Statistics Education Research Journal*, 2(2), 3-46.

Wells, K. (2010). *BP technical briefing and update on Gulf of Mexico oil spil* . Available: http://bp.concerts.com/gom/bptechbriefing051010.htm. Last accessed 15th Feb

Wertheimer, M. (1938). Laws of organization in perceptual forms. In W. D. Ellis (Ed.), *A source book of Gestalt psychology*. London: Routledge & Kegan Paul.

Wickens, C. D. & Carswell, C. M. (1995). The proximity compatibility principle: Its psychological foundation and relevance to display design. *Human Factors*, 37 (3), 473-494.

Wilson T D (1994) The Proper Protocol: Validity and Completeness of Verbal Reports. *Psychological Science*, Vol 5, no 5, pp 249-252

Wilson, T. D., & Schooler, J. W. (1991). Thinking too much: Introspection can reduce the quality of preferences and decisions. *Journal of Personality and Social Psychology*, 60, 181–192.

Winn, N , W . (1987). Charts, graphs and diagrams in educational materials. In D . M . Willows & H . A . Houghton , Eds . *The Psychology of Illustration. Volume 1 Basic Research.* pp . 152 – 198 New York, NY : Springer-Verlag .

Young, R., & Sullivan, P. (1984). Why write? A reconsideration. In R. J. Conners, L. S. Ede, & A. A. Lunsford (Eds.), *Essays on classical rhetoric and modern discourse* (pp. 215–225). Carbondale, IL: Southern Illinois University Press.

Zacks, J., & Tversky, B. (1999). Bars and lines: A study of graphic communication. *Memory and Cognition, 27*(6), 1073–1079.

Zacks, J., Levy, E., Tversky, B., & Schiano, D. J. (1998). Reading bar graphs: Effects of extraneous depth cues and graphical context. *Journal of Experimental Psychology: Applied, 4*, 119–138.

Zajonc, R. B. (1965). Social facilitation. *Science, 149*, 269–274.

Zhang, J. (1966). A representational analysis of relational information displays. *International Journal of Human-Computer Studies*, 45, 59-74.

Appendix

Item 1 (verbal protocol transcript from Experiment 1, line graph condition).

Trial 1 Graph 3

Reads title

I don't know what cloud seeding — Ignoring x variable

When there's high rainfall there is cloud seeding

When there's less rainfall there'll be lower cloud seeding

Trial 2 Graph 6

Reads title

Beef has higher ….

Identifies x1 x2 Y — Content specific error (graph six)

Beef is good for weight gain, cereal isn't

Beef has high and low protein type and cereal doesn't, so beef is better for weight gain.

Trial 3 Graph 2

Reads title — (Miscellaneous)

When it's cold there's low stress when it's hot there's high stress

Fractures are more common in hot temperatures and more average in cold

Trial 4 Graph 4

Reads title (Correct response – elementary)

Higher exercise high well being than low exercise same for males and females

Trial 5 Graph 1

Reads title

Words task Aa lower RT than pictures (Correct response – elementary)

Task Ab RT quicker for words than pictures (opposite)– a considerable difference

Trial 6 Graph 5

Reads title (Correct response – elementary)

% error is the same whether low or high experience during the day

At night much higher at low experience than high – considerably