UNDERSTANDING GRAPHS WITH TWO INDEPENDENT VARIABLES

by

Jennifer L. Cooper

A Dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Psychology

written under the direction of

Dr. Rochel Gelman

and approved by

________________________

________________________

________________________

________________________

New Brunswick, New Jersey

October, 2010
Adults are not necessarily competent users of graphs with two independent variables, despite the frequency of this representational format. The three tasks in this thesis address the impact of interpretation statements and graph patterns. Interpretation statements were based on the statistical effects – simple effects, main effects, and interactions. Graph patterns were systematically varied based on a novel classification scheme of graphs with two IVs. I suggest that the complexity of a graph’s data pattern depends on the consistency of the simple effects’ directions and magnitudes.

In the first study, undergraduates constructed graphs based on statements about data patterns. Errors reflected a misunderstanding of how two IVs could be combined and represented graphically. When the experimental group had graph-relevant information added (variable labels spatially located on axes), the ability to represent the relationships among the IVs significantly increased. The ability to satisfy the constraints imposed by the statements was not affected. Adding labels specifically
targeted skills relevant to graphical literacy. Transfer to a third trial was stronger for those of higher math abilities.

The second study focused on the effect of an introductory statistics course. Overall, undergraduates performed well on statements describing the simple effects of the IVs. However, even though they improved from Time 1 to Time 2 for interaction statements, performance on statements about main effects and interactions still showed considerable room for improvement.

In the third study, repeated trials of the 20 patterns proposed by the simple effects consistency model established that the proposed classification scheme addresses additional sources of variability in reasoning with graphs (i.e., sources not captured by traditional classification schemes). As the complexity level of the data pattern increased, performance (based on accuracy and RT) decreased, with parallel impacts on performance for each IV’s complexity. This suggests that participants responded to conceptual differences among the levels, as the graph’s perceptual characteristics vary based on the IV.

Further development of a model organizing graph patterns will allow investigation of the interplay between the statement and graph pattern. In turn, this can lead to greater understanding of the graphical reasoning processes and improvements in graphical literacy.
Acknowledgements

This thesis would not have been possible without the support of numerous people in my life. I am grateful for the support of my advisor, Rochel Gelman, and the members of my committee: Gretchen Chapman, Lance Rips, and Manish Singh. My colleagues, friends, and family provided immeasurable support during this time.
Chapter 1: General Background about Graphs

1.1 Overview
1.2 Specific Aims
1.3 General Background about Graphs
   1.3.1 Graphs as External Representations
   1.3.2 A Note on Terminology
   1.3.3 Viewing Graphs in Context
1.4 Contexts and Formats of Encountered Graphs
1.5 Representational Choices with Graphs
   1.5.1 Misleading or ‘Bad’ Representational Choices
   1.5.2 Making ‘Good’ Representational Choices
1.6 Graph Interpretation: Effects on Performance
   1.6.1 Graph Characteristics
      1.6.1.1 The case of lines or bars
2.5 Categorizing Graph Patterns ................................................................. 60

2.5.1 Classification Based on the Main Effects and Interactions

(Traditional Classification) ................................................................. 61

2.5.2 Variations on the Traditional Scheme ............................................ 63

2.5.3 Classification Based on the Consistency of Simple Effects .......... 65

2.5.3.1 Overview of the simple effects model ...................................... 67

2.5.3.2 Levels of simple effects consistency ..................................... 69

2.5.3.3 Graphical representation of each level of simple effects

consistency ...................................................................................... 73

2.5.3.4 The simple effects consistency model for both IVs .......... 77

2.6 Comparing Graph Classification Schemes ........................................... 81

2.7 Assumptions .................................................................................... 83

2.7.1 Geometric Patterns ..................................................................... 85

CHAPTER 3 ................................................................................................. 89

OVERVIEW OF EXPERIMENTAL TASKS ............................................. 89

3.1 Participants ..................................................................................... 91

3.2 Tasks ............................................................................................... 92

CHAPTER 4 ................................................................................................. 94

DRAW THE DATA: ADDING GRAPH-RELEVANT INFORMATION .......... 94

4.1 Overview ......................................................................................... 94
CHAPTER 6................................................................................................................... 182

SPEEDED EVALUATION OF GRAPH-STATEMENT MATCHES........................................... 182

6.1 Overview................................................................................................... 182

6.2 Stimuli....................................................................................................... 184

6.2.1 Context Domains ................................................................. 184

6.2.2 Statements .............................................................................. 188

6.2.3 Graph Stimuli........................................................................... 194

6.2.3.1 Constraints and procedures for generating data values. 195

6.2.3.2 Using the simple effects consistency model to generate
data values............................................................................ 199

6.3 Overall Design of Statement – Graph Pairings ........................................... 203

6.4 Procedure.......................................................................................... 205

6.4.1 Timing.......................................................................................... 209

6.5 Analyses Plan Based on Experiment Design................................................ 210

6.5.1 Summary of Design of Analyses .................................................. 212

6.6 Participants ....................................................................................... 213

6.7 Trial Validity and Missing Data.............................................................. 214

6.7.1 Invalid Trials based on RT ...................................................... 215

6.7.2 Invalid Trials Based on Response............................................. 218
6.7.3 Excluding Subjects and Working with Missing Data ...................... 218

6.8 Results: Overall Performance ................................................................................................. 220

6.9 Results: Effects of Statement ................................................................................................. 222
   6.9.1 Effects of Statement on RT ................................................................................................. 222
   6.9.2 Effects of Statement on Accuracy ...................................................................................... 226

6.10 Results: Effects of Graph Pattern .......................................................................................... 229
   6.10.1 Effects of Graph Pattern on RT ......................................................................................... 229
   6.10.2 Effects of Graph Pattern on Accuracy .............................................................................. 230

6.11 Results: Influence of Math & Statistics Background .......................................................... 233

6.12 Results: Interactions Among Statements and Graphs ....................................................... 235
   6.12.1 Interactions between Statement and Graph on RT ......................................................... 235
   6.12.2 Interactions between Statement and Graph on Accuracy ............................................. 236

6.13 Combining and Summarizing Accuracy and RT Measures .............................................. 238
   6.13.1 Effects of Statement ........................................................................................................... 239
   6.13.2 Effects of Graph Pattern ..................................................................................................... 240

6.14 Are these patterns present in the individual participants? ................................................. 241

6.15 Summary and Discussion ..................................................................................................... 243

CHAPTER 7 ...................................................................................................................................... 247

GENERAL DISCUSSION .................................................................................................................. 247

7.1 Assessing the Simple Effects Consistency Model: Effects of Graph Pattern on
   Graphical Literacy ....................................................................................................................... 248
7.1.1  A Comment on Categorizing Graph Patterns ......................... 250

7.2 Effects of Interpretation Statements on Performance .................. 251
   7.2.1 Types of Interpretation Tasks .............................................. 252
   7.2.2 Other Types of Interpretation Tasks ..................................... 254

7.3 Graph Construction ................................................................. 256

7.4 Influence of Math Background .................................................. 258

7.5 A Comment on Interpreting Interactions .................................... 259

7.6 Future Directions ........................................................................ 261

Appendix 1 ......................................................................................... 264

Descriptive Statistics from Entire Sample in Speeded Evaluation of Graph-Statement Matches ............................................................. 264
Effects of Statement on RT ............................................................... 264
Effects of Graph Pattern on RT ......................................................... 265
Effects of Statement on Accuracy ..................................................... 267
Effects of Graph Pattern on Accuracy .............................................. 268

Appendix 2 ......................................................................................... 270

Effect of Graph Pattern on Accuracy and RT for Each Statement .................... 270

References ....................................................................................... 281

CV .................................................................................................. 287
List of Tables

Table 2.1  Graph Classification Schemes .................................................................. 61
Table 2.2  Complexity levels based on the sign and magnitude of the simple
effects ................................................................................................................ 71
Table 2.3  Conceptual and perceptual description of each complexity level in the
hypothesized graph space based on the consistency of simple effects .......... 76
Table 2.4  Complexity levels based on the existence of main effects and
interactions ....................................................................................................... 81
Table 4.1  Statements presented to participants about the data pattern which
they were to graph .......................................................................................... 104
Table 4.2  Distribution of graph patterns across subject groups ......................... 107
Table 4.3  Frequencies of reading, writing, and revision (labels & data values)
across participants .......................................................................................... 121
Table 4.4  Revision and effort across trials ............................................................. 123
Table 4.5  Bar graphs predominated as response types ....................................... 128
Table 4.6  Frequency of tick mark usage and unit spacing on the y-axis .......... 134
Table 4.7  Proportion of relationships represented, regardless of accuracy ...... 138
Table 4.8  Proportion of relationships represented accurately (out of all 6
relationships) ................................................................................................. 141
Table 4.9  Math level and statement type both impacted the accuracy of
performance ..................................................................................................... 144
Table 4.10 Performance varied based on the graph pattern ................................. 150

xiii
Table 5.1  Six statements were presented to participants at each time in this task ................................................................. 158

Table 6.1  Independent and Dependent Variables used in the four context domains ................................................................. 190

Table 6.2  Statement types presented as stimuli .................................................. 195

Table 6.3  Correct responses based on statement type ................................. 207

Table 6.4  Characteristics of trials Included in analyses ............................ 217

Table 6.5  Rates of problematic trials in the data set .................................... 218

Table 6.6  Mean accuracy levels by statement and graph pattern set .......... 231

Table 6.7  Reaction time based on graph pattern (across all statements) .... 233

Table 6.8  Statistical results for the main effects of complexity for each IV .... 234

Table 6.9  Mean performance levels by the complexity level of each IV ........ 239

Table 6.10 Accuracy levels based on graph pattern for each graph from the simple effects consistency model ................................. 235

Table 6.11 Inferential statistics for IV Complexity x Statement interactions. .... 240

Table 6.12 Effects of complexity level by statement (IV, directionality, and statement type) ................................................................. 241
## List of Figures

| Figure 1.1 | Example graph stimulus from the current research | 2 |
| Figure 1.2 | Example of a context-free graph used in research | 10 |
| Figure 1.3 | Illustrative example of graph used to assess the communication of graph contents and spoken language skills in English-language learners | 11 |
| Figure 1.4 | Misleading graph based on distortion of y-axis scale | 15 |
| Figure 1.5 | Example line graph displaying an interaction despite a nominal horizontal axis | 19 |
| Figure 1.6 | Effect of switching the locations of IV<sub>A</sub> and IV<sub>B</sub> on the perceptually displayed pattern | 24 |
| Figure 1.7 | Table look-up (based on processes from Cleveland, 1994) | 27 |
| Figure 1.8 | Example of typical graph interpretation tasks along with performance levels | 32 |
| Figure 1.9 | Sample stimulus geography graph with three IVs from Mautone & Mayer (2007, p 642) | 34 |
| Figure 2.1 | Graph depicting interactions of common IVs (humidity and temperature) | 43 |
| Figure 2.2 | Multiple representational formats of an interaction between two independent variables | 45 |
| Figure 2.3 | Graphical representations of 2 IVs that do not interact | 46 |
Figure 2.4  Pattern between two familiar IVs (Type of Bike, Breakfast Location) that was discoverable in Wilhelm & Beishuizen’s (2003) self-directed inductive learning task ................................................................. 52

Figure 2.5  Effects (main effects and interactions) reported by experts and undergraduates based on experimental data they generated (based on data from Schunn & Anderson, 1999) ........................................ 55

Figure 2.6  Traditional classification of graphs with 2 IVs based on the number of statistical effects ........................................................................................................... 62

Figure 2.7  Simple effects depicted on a graph ................................................................. 67

Figure 2.8  Relationships between simple effect signs and magnitudes for one IV differentiate the complexity levels in developed model of graph classifications ................................................................. 69

Figure 2.9  Graphical representation of changes in IVA’s complexity level .............. 73

Figure 2.10 Graphical representation of changes in IVB’s complexity level .............. 75

Figure 2.11 Categorization scheme based on the consistency of simple effects (line graphs) ................................................................................................................ 79

Figure 2.12 Categorization scheme based on the consistency of simple effects (bar graphs) ................................................................................................................ 80

Figure 2.13 Overlap among traditional and simple effects consistency classification schemes .............................................................................................................. 82

Figure 2.14 Prototypical representations in the graph classifications can be endpoints on a continuous dimension of variations in the graph pattern .................. 84
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.15</td>
<td>Geometric patterns applicable to creating a ‘family’ of graphs</td>
<td>86</td>
</tr>
<tr>
<td>2.16</td>
<td>A family of graphs, based on geometric relationships</td>
<td>87</td>
</tr>
<tr>
<td>3.1</td>
<td>Overview of experimental tasks in thesis</td>
<td>91</td>
</tr>
<tr>
<td>4.1</td>
<td>Example of trial in graph construction task</td>
<td>99</td>
</tr>
<tr>
<td>4.2</td>
<td>Graph patterns described by the sets of statements presented to participants</td>
<td>107</td>
</tr>
<tr>
<td>4.3</td>
<td>Experimental design for Draw the Data task</td>
<td>109</td>
</tr>
<tr>
<td>4.4</td>
<td>Added Labels manipulation</td>
<td>110</td>
</tr>
<tr>
<td>4.5</td>
<td>Math levels in the Control and Added Labels conditions</td>
<td>113</td>
</tr>
<tr>
<td>4.6</td>
<td>Constructed graphs depicting the dissociability of the existence and accuracy of the relationship among variables</td>
<td>117</td>
</tr>
<tr>
<td>4.7</td>
<td>Reading marks increased in the Control condition, but not in the Added Labels condition</td>
<td>122</td>
</tr>
<tr>
<td>4.8</td>
<td>Effects of trial and math level on the variety of effort categories shown</td>
<td>124</td>
</tr>
<tr>
<td>4.9</td>
<td>Examples of typical graphs</td>
<td>126</td>
</tr>
<tr>
<td>4.10</td>
<td>Examples of atypical graphs</td>
<td>127</td>
</tr>
<tr>
<td>4.11</td>
<td>Traditional graphs were more frequently represented during and after the manipulation in the Added Labels condition</td>
<td>131</td>
</tr>
<tr>
<td>4.12</td>
<td>Examples of constructed graphs in which 1 IV was placed on each axis</td>
<td>132</td>
</tr>
<tr>
<td>4.13</td>
<td>Example of a non-integrated graph</td>
<td>133</td>
</tr>
</tbody>
</table>
Figure 4.14  Graphs representing all, some, or no relationships by trial ................... 137
Figure 4.15  Representing the existence of relationships increased in the Added Labels condition, but not the Control condition .............................................. 139
Figure 4.16  Accuracy of represented relationships across trial and math level........... 142
Figure 5.1  Sample problem from Time 2 material ................................................ 154
Figure 5.2  Examples of data patterns and graphs from Time 1 and Time 2............ 160
Figure 5.3  Depiction of interpretation processes necessary for answering a directional simple effect statement ................................................................. 163
Figure 5.4  Average performance at Time 1 and Time 2, by subject....................... 164
Figure 5.5  Counts of subjects’ standing relative to chance at Time 1 and Time 2... 166
Figure 5.6  Increase and decrease scores, by subject, based on Time 1 performance.............................................................................................................. 167
Figure 5.7  Accuracy based on statement at Time 1 and Time 2 ............................. 168
Figure 5.8  Effects of Math Level and Time on accuracy for each statement............ 172
Figure 5.9  Effect of math on the nondirectional simple effect statement ............... 176
Figure 5.10 Effect of graph pattern on the nondirectional main effect statement for IV_A ........................................................................................................... 177
Figure 5.11 Effect of graph pattern on the nondirectional main effect statement for IV_B ........................................................................................................... 178
Figure 5.12 Effect of graph pattern on the non-directional interaction statement .. 179
Figure 5.13 Interaction between graph pattern and time on the nondirectional interaction statement ................................................................. 180
### List of Boxes

<table>
<thead>
<tr>
<th>Box</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Box 2.1</td>
<td>Alternate definitions of interactions (Keppel &amp; Wickens, 2004)</td>
<td>47</td>
</tr>
<tr>
<td>Box 2.2</td>
<td>Explanation of simple effects</td>
<td>66</td>
</tr>
<tr>
<td>Box 4.1</td>
<td>Sample prompts that were given to 8th grade students for graph construction</td>
<td>96</td>
</tr>
<tr>
<td>Box 4.2</td>
<td>Introductory text read by participants for each of their 3 trials</td>
<td>102</td>
</tr>
<tr>
<td>Box 4.3</td>
<td>Instructions were printed on each trial of the graphing construction task</td>
<td>111</td>
</tr>
<tr>
<td>Box 6.1</td>
<td>Introductory paragraphs for each of the four context domains</td>
<td>188</td>
</tr>
<tr>
<td>Box 6.2</td>
<td>The notational scheme for the independent and dependent variables that participants read about during this study</td>
<td>193</td>
</tr>
<tr>
<td>Box 6.3</td>
<td>Sample statements from one of the content domains</td>
<td>194</td>
</tr>
<tr>
<td>Box 6.4</td>
<td>Instructions for the Speeded Graph-Statement Evaluation task</td>
<td>211</td>
</tr>
</tbody>
</table>
CHAPTER 1

GENERAL BACKGROUND ABOUT GRAPHS

1.1 Overview

Graphs are used extensively to summarize, represent, and communicate quantitative information (Cleveland, 1994; Friel, Curcio, & Bright, 2001). They are essential to data presentation, record-keeping, and analysis in the fields of mathematics and science (Tuftte, 2001; Wainer, 1997). This kind of representational format has become ubiquitous in formal and informal reports in academics, news media (print and television), museums, and so forth. Little is known about adults’ literacy with graph formats, particularly those involving multiple independent variables.

Graphical literacy refers to the abilities involved in accurately deriving, understanding, and reasoning with data presented in graphs (e.g., Cleveland, 1994; Friel et al., 2001). To be graphically literate, graph readers must integrate data with representational formats and conventions. Effectively and efficiently assimilating these components requires mathematical, spatial, and analytical skills as well as other procedural, perceptual, and conceptual abilities (e.g., Cleveland, 1994; Friel et al., 2001; Timmerman, 2002) – these factors are considered in depth in Chapter 2. It is difficult to use a graph effectively without an adequate understanding of the communication medium and the possession of relevant cognitive processes.

The current research addresses undergraduates’ understanding of the content of graphs depicting the possible relationships between two independent variables (IVs),
including interactions. The particular data patterns represented in these graphs are experimentally manipulated, as are the statements about the data in the graphs. Graphs with two IVs are selected because understanding graphs and interactions is a known area of difficulty, whether they are combined together or not (Green, 2007; Peebles & Ali, 2009; Schaeffer, 1976; Wainer & Velleman, 2001; Wilhelm & Beishuizen, 2003). Figure 1.1 shows a prototypical graph from one of the current experiments. However, the emphasis has generally been on these topics in isolation from one another, despite the prevalence of interactions in graphs.

![Figure 1.1. Example graph stimulus from the current research. There are two independent variables: 1) Amount of Sleep and 2) Sugar Intake that affect the dependent variable: Hand Grip Strength.](image)

Undergraduates’ understanding of the data patterns and graphs is approached by presenting both written text and graphic representations. The goal is to address how a participant’s competence varies based on their use of different graph patterns and when making different kinds of interpretations about the data. Across the three
studies, undergraduates are asked to (1) generate graphical representations from written text and (2) evaluate the match between graphical and written representations of the same information. Thus, participants are asked to both construct and interpret graphs. Multiple dependent measures – representational features of the constructed graph, accuracy in generating data patterns in a graph, truth value judgments, sentence completion choices, and response times – are employed. This variation allows a comparison of the effectiveness of different response types and a more comprehensive assessment of graphical literacy.

1.2 Specific Aims

The combination of experiments presented here is meant to yield a multi-faceted view of individuals’ understanding and use of graphically-presented relationships between two IVs. This research on graphical understanding adds to our knowledge about individuals’ abilities to understand relationships between multiple IVs when using graphical representations. In Chapter 2, a theoretical model of possible data patterns is developed. The approach offers a more nuanced classification of data patterns than is generally found. This model was developed and used to generate stimuli for the three studies in order to better assess when and how written and graphical representations influence graphical literacy. This understanding can lead to techniques that enhance the ability to work with graphs that represent two IVs.
Aim 1: Assessing and enhancing undergraduates’ ability to construct graphs showing interactions of two independent variables

While research has tended to focus on the interpretation of graphs, the ability to construct graphs is an important component of graphical literacy. Even among graph interpretation studies, research on how to increase levels of graphical literacy is limited. Discussions often focus on which graph design characteristics aid interpretation (e.g., Kosslyn, 2006; Shah, Mayer, & Hegarty, 1999), but do not address improving interpretive abilities themselves. A notable exception is work by Mautone and Mayer (2007) who incorporated techniques from reading literacy to address the integration of information communicated in a graph.

In the first study, (Chapter 4 – Draw the Data: Adding Graph-Relevant Information to the Representation), participants first read textual descriptions of data patterns and then generated graphs of the data pattern for the 2 IVs from the text. Participants each complete three trials. Simpler versions of this type of task have been presented to middle school students by Mevarech and Kramarsky (1999). They found that the request to construct a graph revealed different levels of graphical literacy among their participants (see Section 4.1). However, graph construction based on data patterns has not been considered in adult populations.

In addition to assessing the competence of undergraduates when faced with constructing graphs with two IVs, the research analyzes different components of graphical literacy. A main goal in the current study is to identify how different conceptual and processing demands on graph construction contribute to individuals’
difficulties; this understanding leads to the potential to improve graphical literacy. Recognizing where individuals have problems will allow the development of targeted instructional techniques. The construction task presents two types of demands: a) to conceptually generate a data pattern from a series of statements and b) to represent the data pattern using a graph. The manipulation in the study reduces the demands posed by the representational aspect of the task. The amount of relevant representational information provided in the graph representation is varied across individuals: half of the subjects receive additional graph-relevant information and structure to the representation (i.e., variable labels). If the planned manipulation of the input addresses a source of difficulty, participants’ performance should improve.

**Aim 2: Comprehensive assessment of the role of the statement kind and data pattern in undergraduates’ understanding of graphs of two independent variables**

In general, graphing research tends to involve ‘simple’ forms of graphs, for example, graphs presented without reference to external phenomena or graphs with one independent variable with two levels (Friel et al., 2001). (Notable exceptions include Mautone and Mayer (2007), Peebles and Ali (2009), and Shah, Mayer, and Hegarty (1999).) In contrast, in the remaining two studies, contextually-based graphs are used, and the relationships between the two IVs are explicitly manipulated in order to assess the impact of the complexity of the data pattern on graph understanding. It is hypothesized that more complex data patterns (as defined by the consistency of the simple effects, see Section 2.5.3) will present greater difficulties for participants. The
emphasis, however, is also on how the combination of different graph patterns with particular statements influences performance. Combining these two components (the graph pattern and the statement kind) enables the recognition of the impact of these variations on graphical literacy. Recognizing this is a step towards further understanding what reasoning processes and strategies individuals implement when faced with graphs with two independent variables.

The second study (*Chapter 5 – Effects of Statistics Education on Graph Interpretation*) follows up on the effect of statement on graph interpretation abilities found when participants constructed graphs from statements. It also provides an assessment of the effect of a statistics course on graph interpretation. In this study, participants evaluated whether statements accurately described the relationship among variables based on a graph. The undergraduates did so before and after their introductory statistics course.

Finally, in the third study (*Chapter 6 – Speeded Evaluation of Graph-Statement Matches*), both the displayed data pattern and the kind of statement are systematically manipulated. Once again, the graphs and relational statements were presented simultaneously, and participants responded to questions asking about main effects, simple effects, and interactions in the graph. Each participant completed a series of trials that presented pairs of different graph data patterns and statements. For example, one of the statements that might accompany the graph shown in Figure 1.1 was “Sleep amount had [a bigger effect on strength] when eating 20 grams of sugar...
daily than when eating 60 grams of sugar daily”. This study offers a more
comprehensive assessment of the role of the data pattern and statement kind.

1.3 General Background about Graphs

I start by covering the circumstances in which individuals regularly encounter
graphs. This leads to a consideration of research on adults’ abilities to interpret graphs
and the factors that influence their use of graphs. A consideration of the theories and
research findings about the effects of graph format, characteristics of the individual as a
graph reader, and task characteristics provides a way to look at graphical literacy in our
society.

1.3.1 Graphs as External Representations

Contrary to popular belief, the information represented in graphs is not
transparent just because graphs are in a ‘perceptual’ format. Accurate and effective
reasoning with graphs is not an automatic skill; although visual, a graph is not a picture
(e.g., Tversky, Kugelmass, & Winter, 1991; Tversky, Zacks, Lee, & Heiser, 2000). While
some information about the graph and/or its data might be readily perceived and
understood through perceptual processes, in general, graph readers¹ must derive
information and interpretations from a graph (Shah et al., 1999). The combination of
the representational format and the informational content of the graph is the message,
or communication, that a graph author conveys to a reader (Shah et al.). In order to

¹ By using the term graph reader, as opposed to viewer, I am stressing the interactive
and constructive activity of a graph reader who is interpreting a graphic representation.
effectively use a graph and understand its content, a graph reader must have the ability to understand the communicative conventions of a graph.

One of the basic ways that graphs communicate data is through the depiction of spatial relationships. To a limited extent, the spatial aspects enable graphs to have a less arbitrary connection to their referents than many other symbolic representations (Tversky et al., 2000). For example, one spatial analogue is increasing numerical quantities and taller bar magnitudes; these share the concept of increasing amounts (Gattis & Holyoak, 1994; Wainer, 1997). Another example is the use of lines that connect intersecting points to reflect trends in data (see section 1.6.1.1). With these spatial analogues, the mapping between the referent and the representation is not entirely arbitrary. However, using these spatial analogues does not occur in isolation; for example, knowledge that axis values increase in equal intervals (see Gattis, 2002) lets graph readers link the numerical magnitude on the axis, the spatial distance of the bar or line, and the quantity being represented.

1.3.2 A Note on Terminology.

With two independent variables (IVs), it is typical to see only one IV depicted on the x-axis. When referring specifically to the IV graphed on the x-axis, this paper will

---

2 See Tversky et al. (1991) for developmental and cross-cultural implications on the directionality of spatial representations (using temporal, preference, and quantitative relationships).

3 Graphs are not limited to one IV on the X-axis, however. For instance, one could choose to fully cross all the levels of the IVs and represent each of the combinations on the x-axis. In addition, other arrangements of DV and IV are possible (e.g., 2 DVs (M and SD) represented on the same x-axis; 3-D graphs; etc.)
use the term $\text{IV}_A$. Terms for the other IV vary: grouping IV, z-axis, parameter, variable in the legend. Kosslyn (2006) refers to both the legend and parameters; Carpenter and Shah (1998) refer to both the z-variable and the parameter of the lines. Here, I refer to it as $\text{IV}_B$; this variable can be thought of as defining the groups that are compared in a line graph and it is represented in the legend. For each of the IVs, the values that the variable can take on will be referred to as the levels of the variable. In the experiments, the IVs in the stimuli will each have two levels: $\text{IV}_A_1$ and $\text{IV}_A_2$ on the x-axis with $\text{IV}_B_1$ and $\text{IV}_B_2$ in the legend. The dependent variable (on the y-axis unless otherwise noted) will be abbreviated as $\text{DV}$.

Thus, in Figure 1.1, $\text{IV}_A$ is ‘Amount of Sleep’; it has two levels – $\text{IV}_A_1$: ‘6 hours’ and $\text{IV}_A_2$: ‘20 hours’. $\text{IV}_B$ is ‘Sugar Intake’. It also has two levels – $\text{IV}_B_1$: ‘20 Grams Daily’ and $\text{IV}_B_2$: ‘60 Grams Daily’. The DV is the hand grip strength.

1.3.3 Viewing Graphs in Context

When individuals encounter graphs in their daily activities, these graphs usually are not presented in isolation or with abstract variables. Nevertheless, this is not how graphs have generally been presented in research tasks. The graphing literature has historically emphasized the perception of simplified graph forms presented without related content information (e.g., Feeney, Hola, Liversedge, Findlay, & Metcalf, 2000; Zacks & Tversky, 1999). Simplified stimuli are oftentimes used in order to focus on perceptual components and limit interference from other graph comprehension processes (Beattie & Jones, 2002). Friel et al. (2001) point out that context was not
regularly included in graphing research until the mid-1990s, despite the prevalence and potential impact of context in graphs outside of experimental tasks. Others have recently taken the approach of using both types of stimuli with the same task; Hurts (2009) initially used context-free graphs and then followed up that study with an experiment based on graphs presented with a naturalistic context. However, the field is changing, as will be seen through later examples.

Context-free graphs (see Figure 1.2) represent purely abstract relationships that do not have actual referents in the external world. In actuality though, readers see graphs that are embedded in contexts (see Figure 1.3). The context arises from a variety of sources. Meaningful content, be it measurements, unit sizes, or variables, connects the representation to phenomena in the external world is part of a graph’s context.

Figure 1.2. Example of a context-free graph used in research. (Figure from Feeney et al., 2000, p. 154)

Additional context relating to the variables and their referents comes from the use of titles and captions as shown in Figure 1.3. The particular content (e.g., work hours in Figure 1.3) of the context also has a role; the effects of prior knowledge and graph content are considered later (see Section 1.6.3).
The graph below shows what people of two age groups value about their work. Describe the information given in the graph.

**WHAT PEOPLE VALUE ABOUT WORK**

**Figure 1.3.** Illustrative example of graph used to assess the communication of graph contents and spoken language skills in English-language learners. (Figure from Katz, Xi, Kim, & Cheng, 2002, p. 157). The graph has meaningful referents and as such, has context. The same notation introduced in Section 1.3.2 can be used to refer to graphs with more levels for the IVs. Here, $I_{VA}$ is age; it has two levels: Ages 20-30 and Age 50-60. $I_{VB}$, ‘Things Valued About Work’ is not explicitly labeled as such, but it encompasses the five levels listed in the legend. The DV is the % of respondents (although, again, it is unlabeled).

### 1.4 Contexts and Formats of Encountered Graphs

There is wide variability in the types and complexities of graphs that adults regularly encounter. Line and bar graphs predominate across a variety of sources (Kosslyn, 2006). College textbooks are one important source to consider, especially as college students compose the population for much of the research done on adults’ graphical literacy. Around 60% of the graphs in psychology textbooks (introductory and upper level) are line graphs; just under another quarter of the graphs are bar graphs.
Although there is wide variation across psychology textbooks, on average, there is 1 graph for every 10 pages (Peden & Hausmann). These relative and absolute frequencies are similar in other, more general sources as well. Newpapers and magazines in the Zacks, Levy, Tversky, and Schiano (2002) sample primarily used line and bar graphs; Schield (2006) found that bar charts were the predominant format in USA Today. Similarly, line graphs and bar graphs (albeit to a lesser degree) were the most frequent formats found in counts of graphs in journal articles (Best, Smith, & Stubbs, 2001; Zacks et al., 2002).

In addition to this information on the frequencies of the different formats, various surveys have also considered how graph format varies with respect to content area. Within introductory psychology textbooks, Nevid and Forlenza (2005) found graphs to be more frequent in the ‘softer’ chapters (abnormal, social, and developmental). Of his two conjectures regarding this difference, one deserved particular note. This is the possibility that the experimental findings and statistical representations (e.g., d’ or log transformations) in cognition or perception are deemed too complex to be graphed for these students. The discrepancy is of greater concern given the opposite finding in journal articles and content-specific textbooks; in these

---

4 This worked out to an average of 42 line graphs and 15 bar graphs (range: 1 – 41) per introductory textbook (upper level textbooks averaged 26 line and 8 bar graphs). Scatterplots were relatively more frequent in upper level learning and perception texts; overall, content domain affected graph frequency and type (Peden & Hausmann, 2000).

5 Nevid and Forlenza’s (2000) alternative conjecture, that figures and diagrams in the ‘hard’ chapters preclude the use of graphs cannot be a full explanation.
sources, graphs occur more in the ‘harder’ content areas (Best et al., 2001; Peden & Hausmann, 2000).

Overall, though, these surveys primarily tabulate the type and frequency of graphs. Analyses about the other dimensions of graph displays are much less common. Zacks et al. (2002) did delve deeper into tabulating the graphs in their sample, assessing properties such as how many levels an IV had and whether or not a legend was present. Zacks et al. also found that roughly 30-50% (depending on media type) of the graphs they studied had axes or legends that were shared across graphs (e.g., multi-panel graphs or multiple DVs on a single graph) in order to increase the information conveyed by the representation. These latter codings can be used to develop inferences about the presence of multiple IVs in a graph – however, they are not absolute. While Zacks et al. (2002) describe the vast majority of their sampled graphs as relating two variables, it is unclear from the examples given whether they mean two IVs or 1 IV and 1 DV. The newspaper graphs pictured in Schield’s (2006) review showed that these graphs predominantly used only one IV. Thus, how often encountered graphs include two or more IVs remains an open question. In general, the classification schemes have not focused on the number of IVs, let alone the pattern of their relationship. 6

1.5 Representational Choices with Graphs

Given the frequent use of graphs to communicate information across sources such as textbooks and newspapers, it seems obvious that it would be desirable to have

---

6 However, as will be presented in section 2.1, multiple and interacting causes are a feature of our world.
‘good’ graphs. The representational properties of graphs serve to bring together a widely disparate class of representations. As with other non-graphical displays such as tables (Hink, Eustace, & Wogalter, 1998; Wainer, 1997), graphs vary in terms of the ease with which they are understood, as well as the particular interpretations readers reach. Graph construction involves specific design choices that affect the communication, that is, choices about formatting affect the interpretation a reader can extract (Reese, 2006). Nevertheless, it is easy to find bad examples of graph design.

1.5.1 Misleading or ‘Bad’ Representational Choices

Many compilations of ‘bad’ or misleading graphs exist, be they in the research literature, the popular press, or online. The inappropriate representations can impair one’s understanding of the content (e.g., Arunachalam, Pei, & Steinbart, 2002; Huff & Geis, 1993; Jones, 2006). Common examples are graphs in which the x-axis decreases from left to right, graphs with scales on the y-axis that do not appropriately match the range of data values, graphs accompanied by insufficient information to determine what the reference value is for percentages (Schield, 2006), and graphs displaying ratios of bar heights that do not match numerical magnitudes (see Figure 1.4) (Arunachalam et al.; Beattie & Jones, 2002). In fact, given the myriad of representational choices, it is probably easier to design a bad graph than to design a good graph (Wainer & Velleman, 2001), especially if one ignores the guidelines on how to design a good graph.
Figure 1.4. Misleading graph based on distortion of y-axis scale. (Figure from Arunchalam et al., 2002, p. 157). Arunchalam et al. modeled their stimuli off graphs from accounting reports as stimuli because of the frequency of errors in published reports. Twenty percent of the reports had measurement distortions (in which the spatial magnitude was not proportionate to the numerosity) greater than 10% of the magnitude (Beattie & Jones, 2002). The figure shown here has a distortion of 200%. Readers’ interpretations are affected by distortions of 10%.

1.5.2 Making ‘Good’ Representational Choices

A fairly large literature describes techniques that enable the ‘best’ data representation (e.g., Cleveland, 1994; Kosslyn, 1993, 2006; Tufte, 2001; Wainer & Velleman, 2001). These texts suggest ways graph authors can maximize communication within the rules and constraints specified by a graph’s representational format. Their topics of consideration cover a wide-range of graph-relevant representational properties: axis placement, tick mark frequency, the use color, etc. In general, the recommendations for graph design are chosen to facilitate graph interpretation; the guidelines are also based on presumed ways to take advantage of how people organize knowledge and our human perceptual systems. The suggestions in these books go beyond the basics of graphing conventions to include theoretical assumptions about
perception and understanding. Research to examine these guidelines is surprisingly limited and/or contradictory.

Although they sometimes present opposing suggestions for graphing design, both Cleveland (1994) and Kosslyn (1993, 2006) justify their choices in terms of the reader’s interaction with a graph, often appealing to known psychological effects. When a graph designer considers the variables and data when constructing a graph, good choices can maximize the information that the graph communicates. Zacks et al. (2002) mention the example of making format selections based on the number of variable levels and length of variable names. These choices, however, are constrained by convention. Subsequently, as Reese (2006) points out, “Good [graph creators] know the rules and break them for a purpose” (p. 86). One example would be using lines to display interactions even when the IV on the x-axis is not continuous (see section 1.6.1.1). In doing so, one needs to be aware of possible misinterpretations, but ‘breaking the rules’ can still increase the information the graph communicates.

Even so, understanding notational conventions is not sufficient when it comes to fully understanding graph axes and values. For instance, even when one knows that labeled tick marks do not necessarily represent all the possible values (Reese, 2006) and that interpolation is used to find intermediate values, those skills are insufficient for graphical literacy. There are procedural aspects to understanding axis labels (Friel et al., 2001), but there are also conceptual aspects. One must understand the abstract relationship between the measurement of an entity and the entity itself. Cleveland (1994) uses the term ‘scale information’ for graph components such as the values and
units on the axes that give content to the visual patterns the graph reader perceives.

Additional steps – conceptual steps – are necessary to link the spatial representation to the content. These steps take the reader from procedurally matching symbols and keys in the representation to working with conceptual relationships among variables.

Applying graphing conventions positively contributes to graphical literacy, but simply following graphing conventions and avoiding misleading graphs is not enough to guarantee that a reader will understand. In other words, the information represented in technically appropriate graphs is still not transparent. Many other aspects influence their interpretation.

1.6 Graph Interpretation: Effects on Performance

A reader’s primary task in graph interpretation is deciphering data patterns in order to obtain information about data from the graph representation. The difficulty of this task is largely influenced by three areas: (a) graph characteristics, (b) task characteristics, and (c) reader characteristics.  

1.6.1 Graph Characteristics

As discussed when considering graphing conventions, choices about representational elements also come into play with ‘good’ graphs. More generally, it is known that the variations in representational format influence the accessibility of

---

7 Other researchers (e.g., Peebles & Cheng, 2003; Shah, Mayer, & Hegarty, 1999) have also employed these three factors as an organization scheme for understanding any type of diagram (graphic) based reasoning.
certain types of information and interpretation, even when the format is technically appropriate (e.g., Shah & Carpenter, 1995; Shah & Hoeffner, 2002; Zacks & Tversky, 1999). Overall, different representational formats emphasize different aspects of the information (Friel et al., 2001; Wainer, 1997).

Graph format is particularly relevant when a graph displays interactions. In these situations, there representational choices that are relevant to illustrating interactions are more numerous. The particular choices of graphing conventions and formatting decisions influence the interpretation of graphs designed to show multiple variables. Two such cases are considered in detail below: (1) lines versus bars and (2) variable assignment to locations. Both of these cases make use of the ease of perceiving trends and groupings.

1.6.1.1 The case of lines or bars

At some point, most individuals have probably heard the general guideline that bars should be used for discrete IV and lines should be used for continuous IV (as reported in Tversky et al., 2000 and many statistics texts). However, this representational choice is more nuanced than what the approach based on the variable’s scale would suggest.

A second approach to choosing lines or bars as the format focuses on the desired interpretation. Kosslyn (2007) strongly recommends the use of line graphs to display interactions (see Figure 1.5) and the use of bar graphs when the point values are needed. Peebles and Ali (2009) specifically refer to these types of graphs as ‘interaction
graphs’ when two IVs, each with two levels, are graphed using lines. The lines are beneficial specifically because they form a familiar (and more easily perceivable or recognizable) pattern (Reese, 2006; Kosslyn). For Kosslyn (and others) this focus on interaction patterns trumps the legality issue of a continuous or discrete measure. Kosslyn makes this recommendation even when the IV on the x-axis has a nominal scale. His message is that lines are to be avoided only when they would result in a misinterpretation.

![Figure 1.5. Example line graph displaying an interaction despite a nominal horizontal axis. As Kosslyn (1993, 2007) conveys, line graphs offer an easier interpretation of the interacting pattern.](image)

Tversky et al. (2000) suggest a third approach to choosing between lines or bars based on the spatial properties of the symbolic forms in graphs. As general symbols,

---

8 It is possible (and perhaps likely) that such patterns are useful only for experienced graph readers (Peebles & Ali, 2009). However, then the question remains of how to define an experienced graph reader.
bars act as containers and lines act as connectors – they consider these inherent visual properties of the symbols. Then, in their more constrained role as meaningful graphic forms, these symbols gain further meaning. Bars signify the piling up of discrete items; lines connect and emphasize trends. Thus, the approach of Tversky et al. also focuses on the influence of the format on the graph’s communication, but the rationale is different from Kosslyn’s approach. Tversky et al. rely on the inherent visual properties of the symbols as a means of negotiating the different graph formats.

Zacks and Tversky (1999) compared the foregoing spatial / symbolic approach to symbol meaning to a discrete versus continuous approach by presenting adults with simplified line and bar graphs of the same data. In one case, age (2 levels) was selected as a continuous IV; in the other case, gender was selected as a discrete IV. The continuous DV of height was the same for both cases. With line graphs, participants produced interpretations that mentioned trends (e.g., as someone gets older, they get taller). With bar graphs, interpretations tended to be about discrete comparisons (e.g., boys were taller than girls). These co-occurrence patterns (line – trend; bar – comparison) replicated on a graph production task. This dichotomy between trends and discrete comparative statements had a greater effect on performance than did the nature of the variable (i.e., discrete or continuous IV).

Peebles and Ali (2009) classified participants’ verbal reports as they interpreted graphs and identified the types of interpretive tasks their subjects performed with bar graphs.

---

9 Although rare, some college students did incorrectly treat gender as a continuous dimension when it was plotted on the x-axis of a line graph (Zacks & Tversky, 1999).
and line graphs of the same data. Participants reading bar graphs generally were able to make comparisons among data, but only half of the line graph readers were at this level. Of the remaining line graph readers, a substantial portion of the sample had difficulty with determining the magnitude represented at a particular value for the x-axis (this is point-reading: Friel et al.’s 1st level of graph interpretation (see Section 1.6.2)). Peebles and Ali (2009) extrapolated from this that line graphs were more difficult for their participants to process and understand. One caveat to this approach is that participants’ interpretations represent their favored approach to graph reading and not necessarily their highest competence since the responses came from an open-ended test format.

In practice, then, Tversky et al.’s (2000) approach linking interpretation type and format can correspond to Kosslyn’s (2007) guidelines for lines for interactions and bars for discrete comparisons. Given the findings of Peebles and Ali (2009), however, selecting between using lines and bars to represent data does not appear to be a straightforward matter – reader and task characteristics play a role. In addition, the difficulty of the interpretation task interacts with the graph format. Shah et al. (1999) remind us that lines enable easier detection of trends; Shah & Hoeffner (2002) review other examples in which format interacts with task. Thus, understanding and performance levels cannot be considered without considering the effects of the graph, task, and the reader.

---

10 Peebles and Ali (2009) note that this subgroup was not limited to the undergraduate students; it also included graduate students as well as academic and non-academic staff.
1.6.1.2 The case of variable placement

Another decision arising with graphs with multiple variables is how to assign variables to their location on a graph. In terms of graphing conventions, Kosslyn (2006) suggests that the most important IV (that which is the focus of the primary contrast of interest) belongs on the horizontal axis; Shah, Mayer, and Hegarty (1999) concur based on their research findings (discussed below). Kosslyn suggests that the other IV should then be treated as parameters and marked with symbols through the legend and/or labels. If, however, the IVs are of equal importance to the data, other characteristics of the IV, such as scale or number of levels, become relevant. Kosslyn suggests that interval variables (with more levels) be assigned to the horizontal axis. Pattern simplicity is a subsequent approach to determining variable placement. These procedures and suggested conventions for assigning variables to graph locations are based on a desire for the graph organization to match the intended message or interpretation.

The emphasis on the effect of variable placement on interpretation is supported by the empirical research. The formatting decision arises less frequently when there is only one IV. With only 1 IV, the IV is usually on the horizontal axis with the DV on the vertical axis (Kosslyn, 2006; Tversky et al., 1991). However, Gattis and Holyoak (1994) found that placing the IV on the vertical axis and the DV on the horizontal axis led to better understanding of the causal relationship represented by the graph. They suggest that this enabled participants to consider the cause before the effect. Varying the
placement of the IV (however many IVs there are) results in differential ease of causal interpretations of informationally equivalent graphs.

Shah et al. (1999) worked with social studies graphs in which the graph creators had originally wanted to emphasize trends. When reading the original graphs, participants tended to summarize successive table look-up operations. They simply reported the values for each level of the variables. Shah et al. modified the graphs by changing variable location and perceptual groupings; these modifications result in different visual chunks. Subsequently, the content of the interpretations changed, as measured both by statement verification tasks and verbal descriptions of graph content. The modified versions had more reports of the trend than the original versions.

Participants made use of which quantities were represented adjacent to each other or connected by lines. These visual chunks that were accessible from the graph influenced the type of interpretation that participants made.

Studies investigating the issue of variable placement have presented the same data set in graphs that only differ in where each IV is located; the IV on the x-axis and the IV in the legend switch locations (as shown in Figure 1.6). Usually, interpretations do not depend on the specific content of the IV as much as they depend on the location of the IV – whether it is on the x-axis or on the legend. In accordance with this finding, Peebles and Ali (2009) collapsed participants’ responding across the two perspectives by cataloguing responses in terms of IV_A and IV_B rather than the specific content of the variable.
Figure 1.6. Effect of switching the locations of IV<sub>A</sub> and IV<sub>B</sub> on the perceptually displayed pattern. On both graphs, the hypothetical data pattern is one in which computer type has no impact on crashes until the computer is older. Not only do crashes increase more in older computers, they increase by a greater amount for desktops than for laptops. The data to support this interpretation is shown in both graphs; however, which aspect of the interpretation graph readers focus on depends on which variable is on the x-axis.

When an IV was located on the x-axis, participants tend to interpret the relationship between the IV and the DV (Carpenter & Shah, 1998). In contrast, participants tended to use the variable in the legend (IV<sub>B</sub>) in a nominal fashion. To understand the quantitative relationship between the levels of IV<sub>B</sub> and the DV requires comparing across the two levels of IV<sub>B</sub> (i.e., across lines), rather than comparing values within a single line. Carpenter and Shah note that making the comparison across lines requires more inferential reasoning, and thus, understanding the quantitative relationship between the levels of IV<sub>B</sub> and the DV is more difficult. Additionally, Carpenter and Shah attribute the greater understanding of IV<sub>A</sub> and DV (compared to IV<sub>B</sub> and DV) to the same reason that lines are preferred over bars for interaction patterns: the salient trend and pattern information.
In summary, the choice of line or bar and the effect of variable placement capture a main finding in the literature. This is that the interpretation of a graph is not independent of its representation. Although the actual data and content are independent of the chosen representation (Tversky et al., 1991), the same data can be represented in variety of formats. Still, the choice of representational elements lessens the data independence as the data becomes part of a representation. While it is theoretically possible to obtain all the information from the representation, the particular representation constrains the interpretations that are generated by graph readers. It is the graph reader’s responsibility to work to decipher the independence and obtain the data and interpretation from the representation.

Overall, the empirical data suggesting more nuanced consideration of the guidelines converge on two points: the principle that the goal of good graph design is to reduce the cognitive work of the graph reader and the general finding that representational format influences interpretation. As seen in the two cases, the formatting decisions can influence interpretation. The graph reader then must link the physical and spatial information presented in the representation to the data values and labels in order to reach an interpretation of the data (Cleveland, 1994; Friel et al., 2001).

---

11 Shah et al. (1999) provide an additional clear example using another type of formatting decision; they represented the same data using two different numerical representations. The type of values used on the vertical axis influenced the type of interpretive statements that were more accessible. Undergraduates’ accuracy on relative comparison questions was higher when the scale had used percentages; recalling absolute trends and facts was relatively easier when the graph’s scale had used absolute frequencies.
1.6.2 Task Characteristics

Theories concerning the graphing task’s characteristics focus on the interpretive processes of graph use. In general, the more difficult tasks are those requiring the reader to integrate the data information and representation into a cohesive whole that makes use of both the representation and an individual’s knowledge. The more complex tasks involve broader analyses that integrate more components of the data, representation, and an individual’s knowledge. In contrast, simpler tasks rely only upon local features (i.e., features at the representational level) of the graph. According to Friel et al. (2001), these interpretation tasks can be divided into the following three levels.

The first (and simplest) interpretation level Friel et al. (2001) outline is reading and recognizing specific data values on the graph, a process referred to as table look-up or point-reading. Cleveland (1994) decomposes this process into three steps: scanning, interpolation, and matching (see Figure 1.7).
Q: How many words were remembered after 3 minutes of studying in a noisy environment?
A: 5.5 words.
- **Scan** horizontal axis for the X-value of 3 minutes.
- **Interpolate** if necessary.
- **Match** the light colored bar to the noisy group.
- **Scan** from the height of the bar at X = 3 to the value on the y-axis. **Interpolate** if necessary.

**Figure 1.7.** Table look-up (based on processes from Cleveland, 1994). Table look-up or point-reading arises from a series of processes in which individuals determine the value of a magnitude represented on a graph.

In all three of these steps, the content-based scale information is referenced.

Together, these processes take the graph user from the requested conceptual information to the axis values. First, scanning involves finding the value on the axis.

Then, interpolation is used if the value is not physically there, and the graph reader must locate it between the labeled values on the axis. Additionally, when legends are
involved, the reader must match the graphing notations with the conceptual information. These sub-processes accomplish table look-up, which is also referred to as ‘point reading’; both terms reflect the underlying operations and the reliance on the physical representation of the graph. Graph users operating at this task level are essentially encoding or decoding information (Konold & Khalil, 2003).

At this first level of interpretation, reading data values from a graph requires knowledge of graphing conventions. The explanation of the steps above makes reference to the axes with their tick marks and labels as well as the legend. In addition, table look-up using a graph also requires the recognition that the graphical representation stands in for the data.

The more advanced interpretation tasks require more than one step to reach an answer. The second level of interpretation theorized by Friel et al. (2001) considers computations or comparisons of individual data points; an individual at this level is ‘reading between the data’ and performing some type of manipulation on the data that they gather. This level of interpretation task requires the first level (i.e., table look-up) as well as the subsequent use of that information. For instance, the second level of interpretation might address how many categories exceed a certain frequency, given by either the graph or an additional source. Pattern perception (Cleveland, 1994) is employed as readers use the information of the bar heights to make rankings or comparisons. At this level, the reader may be attending to more than one category. Graph usage at this level is very common (Zacks et al., 1998).
The **third** level, the most advanced level of interpretation Friel et al. (2001) propose, looks at trends or predictions. As such, it is based on more of the graph's information. Additional contextual information is brought to the interpretation tasks. Just as the second level of interpretation also included more information, there is a continuing integration of multiple steps and the use of general features about the graph as well as its context (e.g., prior knowledge about the content domain). At this level of interpretation, the reader must attend to more than one category and data value in each response.

Thus, in summary, steps in interpretive tasks can be rank-ordered - moving from reading data values, to interpreting the patterns and trends, and then finally to a contextually-based purpose or overall understanding of the graph. Each successive level adds complexity to the interpretive task and increasingly requires the application of additional knowledge – knowledge about both the content of the graph and how the content relates to the representation. As the level of interpretation increases, the reader takes away more information about the data. One needs to work at multiple levels of interpretation in order to be able to gather all of the information that the graph communicates.

**1.6.2.1 Performance and assessment at different interpretation levels**

In general, the above ranking of interpretation levels is empirically supported (e.g., Doig & Groves, 1999; Parush, Ferres, Rasouli, & Lindgaard, 2006). Tasks of lower complexity levels such as table look-up have higher accuracy levels and faster reaction
In addition, there is some suggestion that an earlier interpretation level is needed – that is, a level that does not presume competence with point-reading. The interpretive levels have also been used as a way of ranking individual graph readers. Peebles and Ali (2009) used Friel et al.’s (2001) three levels of interpretation tasks to categorize the verbal think-aloud statements of the graph readers in their study and then to determine the level at which participants were operating. However, they also found it necessary to add an earlier level reflecting difficulties that their adult subjects showed with point-reading.

While there is sufficient research evidence to rank the difficulty of the interpretive tasks, the majority of the research focuses on less complex tasks. Only a small portion (around 10%) of the questions in one assessment of high-schoolers were open-ended narrative responses (e.g., what is the relationship between two variables?). The vast majority focused on point-reading or comparison between two values—tasks at the first and second levels of interpretation (Aberg-Bengtsson & Ottosson, 2006).

In a normative assessment of undergraduate’s knowledge after a statistics course, almost half of the questions involved graphs – generally histograms (Garfield,

---

12 In terms of reaction times, Hink et al. (1998) found their participants answered numerical questions (~ 40 sec) faster than questions about trends or comparisons (~ 50 sec) on line and bar graphs (p. 444). This increase in time is appropriate because the latter two question types both require more than one table look-up operation as well as an action upon the obtained values.
delMas, Chance, & Ooms, 2006). Students averaged 62% (SD = 20%) correct on the graph problems. While the problem types reflect the course material, it is notable that the graphing questions primarily asked students to match descriptive statistics and interpretive statements to the graphed data distributions. This places the questions at Friel et al.'s (2001) 1st and 2nd levels of interpretation. Despite extensive coverage of histograms in a first course in statistics, difficulties working with the graphs persisted.

In general, regardless of the format of the graph, assessment tasks involving higher-level graph interpretation skills are less common (Konold & Khalil, 2003). Even when assigned tasks could be taken to more complex levels, assessment questions seldom ask readers to integrate the information from two or more IVs. For instance, while Hink et al. (1998) used stimuli with two IVs, their interpretation questions relied primarily on table look-up ability with minimal emphasis placed on the interaction and relationship between the two IV. Hink et al.'s (1998) numerical, trend, and comparison questions (see Figure 1.8 for the sample stimuli and performance levels) exemplify typical approaches.
Sample Questions

“(a) numerical, e.g., what was the price of stock 2 during week 5?
(b) trends, e.g., if you bought stock 1 during the first week and sold it during the fourth week, would you have made any money?, and
(c) comparisons, e.g., which stock was less expensive – stock 2 during the third week or stock 1 during the fourth week?”

(Hink et al., 1998, p. 442)

<table>
<thead>
<tr>
<th>Proportion Correct Responses</th>
<th>Hink et al. (1998)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numerical Questions</td>
<td>Trend Questions</td>
</tr>
<tr>
<td>Line graph</td>
<td>.49</td>
</tr>
<tr>
<td>Bar graph</td>
<td>.62</td>
</tr>
</tbody>
</table>

Figure 1.8. Example of typical graph interpretation tasks along with performance levels. Sample task and data from Hink et al. (1998, p. 442), 63 undergraduates, between-subjects design. Viewing bar graphs was helpful with numerical and comparison questions, and the two formats had similar levels of performance for interpretation questions about trends.

The numerical questions involved one table look-up operation; comparison questions required comparing the values obtained in two table look-up operations. While the numerical questions clearly map to point-reading and comparison questions map to the second interpretation level Friel et al. (2001) uses, Hink et al.’s trend questions are not as straightforward. The trend question asked for the comparison of two magnitudes, but it asked about it in the context of the original data, that is, asking participants to link back to the bars’ referents of the two stocks. As other information (e.g., lowering price implies not making money) is involved to some degree, this
question seems to be a mixture of the second and third interpretation levels. Not all interpretation tasks clearly fall into distinct levels in Friel et al.’s hierarchy.

While Hink et al. (1998) found relatively lower accuracy scores on numerical questions, it seems likely that the study underestimated point-reading skills. Other researchers generally have considered point-reading on these types of graphs a skill that adults possess. Lower accuracy in the Hink et al. study on the numerical questions could reflect issues of precision in estimation, given that at least some of the same skills were required to answer the questions at the higher interpretation levels – it was necessary to do at least approximate point-reading in order to answer questions about trends and comparisons. Regardless, these performance levels do not reflect high levels of competence with basic graph comprehension tasks.

Thus, the question is about competence and performance with more difficult interpretation tasks and graphs. Multiple examples of research presenting more complex graphs comes from the research conducted by Mautone, Mayer, Shah, and Carpenter over the past 10 years.

Mautone and Mayer (2007) used written responses to measure undergraduates’ understanding of complex geography graphs, such as that shown in Figure 1.9 about the height of sediment in a river. They measured two types of interpretations: relational statements (specifying the pattern of the variables and their trends in relation to one another) and causal statements (expressing a plausible and relevant mechanism or cause in the functional relationship of the variables). The involvement of multiple
variables, trends, and external knowledge in the task places the reader’s interpretation
tasks at Friel et al.’s (2001) third level.

Figure 1.9. Sample stimulus geography graph with three IVs from Mautone & Mayer
(2007, p. 642). The graph illustrates the effect of three IVs: Time (before or after flood
demarcated in legend), Distance from Bank (continuous variable on X-axis), and Miles
Down the River (represented in separate graphs) on the elevation of sediment in the
river. The horizontal line marks water level.

Mautone and Mayer (2007) implemented various cognitive aids to attempt to
improve students’ interpretations. One manipulation showed participants a similarly
structured graph based on familiar content prior to seeing the geography graph. For
instance, participants saw a graph and interpretation that included the concepts of
driver age, training, and accident rate. The structural relationship between these
familiar variables was identical to that in the geography graph. This manipulation was
intended to prime organizing processes. Relational statements were more frequent (almost 5 statements on average as compared to 3 in the control group [out of 10]) when participants saw a similarly structured graph prior to the target graph. A different manipulation provided students with diagrams and explanations about the scientific processes prior to the presentation of the graph; the manipulation was intended to prime integrating processes. In this condition, students increased the number of causal statements they produced (from an average of .4 (out of 4) in the control group to 1.1 in the experimental group). It is especially noteworthy that the type of manipulation differentially affected the way in which interpretations improved. When integrating processes were primed, causal interpretations increased. When organizational processes were primed, relational statements among the variables increased.

Two main points are important to take away from the foregoing research. First, Mautone and Mayer (2007) successfully improved (with large effect sizes) the students’ interpretations. Second, in the most successful conditions, on average, students still made only half of the possible relational statements and one-quarter of the possible causal statements. Thus, while the improvement shows the effect of prior knowledge as well as graphing skills, the task was still exceedingly difficult for the undergraduates.

1.6.3 Reader Characteristics

In addition to using knowledge of graph design and conventions, graph readers must implement additional conceptual and perceptual skills to complete the interpretive tasks that graphs present (Tversky et al., 1991). The application of these
skills and the interpretation of the data do not develop in isolation; the reader is actively involved in interpretation. Two primary factors individuals apply are general mathematical and graphical skills and knowledge about the phenomena shown in the graph.

### 1.6.3.1 General graphical skills

One of the underlying skills for graphical literacy is knowledge of graphing skills, such as the conventions about the axes discussed in Section 1.5.2. However, these skills are not necessarily enough. Graphical knowledge can relate more generally to mathematical skills. As one example, discrete bar graphs can be interpreted using counting skills while working with line graphs requires a greater understanding of geometrical relationships.

Research is mixed on whether underlying math abilities have been identified as predictors of graphing ability. Aberg-Bengtsson and Ottosson (2006) found that math grades were correlated with a unidimensional ‘general’ factor on a graphing test in Swedish 9th-graders. A possible concern regarding generalization is that their graphing tasks focused on the first and second of Friel et al.’s (2001) levels of interpretation, even though the graph format and interpretation level were not factors affecting performance. However, Shah and Carpenter (1995) also found supporting evidence for an effect of math SAT levels on graph usage. This is not a clearly specified background variable, however, as the SAT includes graph items as well.
Other researchers, however, suggest that math skills are not predictive of performance (Attali & Goldschmidt, 1996). Mautone and Mayer (2007) assessed graphical skill level; while they used this information to ensure that the conditions were equally matched, the skill level did not play into analyses of performance on the graph interpretation task.

At times, difficulties with graphing skills are considered to be predictive of difficulties in other content areas. Michael (2007) found that physiology teachers identified the use of graphs or other mathematical formats as the one of the top 5 reasons that students find physiology a difficult subject.\footnote{Further down the list (9th out of the 18 items considered) was the issue that students would ignore the graphs (Michael, 2007, p. 38).}

1.6.3.2 Previous knowledge

A reader’s previous knowledge can refer to specific information related to the graph content as well as more general underlying principles from the domain. Graph research has ranged from specialized scientific phenomena (e.g., Mautone & Mayer, 2007), to variables selected based on their familiarity (e.g., Peebles & Ali, 2009), and to context-free graphs. The focus on graph research on single domains, contrived situations, or context-free graphs has offered limited insight into the interplay between content knowledge and performance. However, both Aberg-Bengtsson and Ottoson (2006) and Ainley, Pratt, and Nardi (2001) comment that interest and familiarity with the content domain influence graph performance.
Research from other literatures offers further support for the idea that previous knowledge about the content influences performance. The evidence points to the effects of content within graphing with children (Cooper, Brenneman, & Gelman, 2006) and other domains such as experimental design and inductive reasoning (e.g., Lazonder, Wilhelm, Hagemans, 2008; Wilhelm & Beishuizen, 2003) and domain expertise on different levels of scientific comprehension tasks (Royer, Carlo, Dufresne, & Mestre, 199). In many cases, providing familiar content is beneficial.

However, working with familiar content is not always beneficial, and individuals’ previous knowledge can influence other aspects of graph interpretation. For instance, individuals are more likely to look for confirming evidence as opposed to disconfirming evidence (Dunbar, 1993). Thus, it is likely that the degree to which data agrees with a person’s beliefs is one manner in which previous knowledge can influence an individual’s graph interpretation processes and outcomes. Another type of evidence comes from the common finding that new material is easier to learn when it can be anchored within an already existing knowledge base (Hartnett & Gelman, 1998; Royer et al., 1999). This general finding has been demonstrated in the graphing literature: providing additional information about the graph content anchored participants’ interpretations and led to more comprehensive statements about graph content (Mautone & Mayer, 2007). All of these factors interact when it comes to asking what contributes to individuals’ graph reading abilities.
1.7 Written Information in and around Graphs

Graphs are not typically presented in isolation (Cleveland, 1994; Feeney et al., 2000). Axes labels and units; legends or group labels; and titles, captions, and summaries are all textual information that provide information about the content of the graph. In addition to being accompanied by their own text, graphs and the data they represent are also referred in the text they accompany. The text and graph together make up the whole communication (Aberg-Bengtsson & Ottosson, 2006) in which graph readers interpret the representation.

Considering and using the congruence between text and graphs is a critical component of graphical literacy (Feeney et al., 2000; Schield, 2006; Shah, Mayer, & Hegarty, 1999). Overall, the text in a graph title or caption guides the reader – informing them about variables, important points, and perhaps even relationships among variables. When reading graphs, this information is available to graph readers if they choose to make use of it.

1.7.1 Variable Labels

The labels provided for the variables are a critical aspect of the graph representation, especially when considering the link between the data representation and the conceptual content. Both the variable itself, and its levels, can be labeled. The labels connect the quantity represented on the graph to its conceptual referent. Units are a critical component of this as well – conceptualizing the represented measurements is necessary in order to understand the graph’s content.
1.7.2 Titles

In their books on graphing practices, both Cleveland (1994) and Kosslyn (2006) state that good graph titles involve clear identification of the questions the graph can easily answer. The graph titles also should include the relevant information related to the variables and a focus on the big picture. Neither Cleveland nor Kosslyn explicitly state whether titles should be limited to listing the variables involved in the graph. What counts as relevant information is also unspecified, but examples from their books suggest that using nothing but variable names is an option, but not the only option. In research contexts, one possibility is a formulaic approach emphasizing the causal effect of the IV on the DV. Peebles and Ali (2009) adopt this convention (e.g., “Response Time as a function of Stimulus Type and Task”, “Wellbeing as a function of Gender and Exercise”). From their titles, it appears that the variable on the X-axis was the first IV listed in the title.

Graphs in the popular press offer another avenue for examining what titles used in graphs that have commonly been considered poorly designed. A perusal of the graphs from USA Today presented in Schield (2006) shows the pictorial image with graph elements and titles are superimposed on the picture. Although the sample graphs used by Schield (2006) all involve only one independent variable, titles were not confined to listing the variables on the axes. Some made mention of only one of the variables, and others included results as well.

Others have specifically asked the question of what makes a good title. Looking at technical papers, Senda, Sinohara, and Okumura (2004) sought characteristics that
increased reader interest (and understanding). These ‘good’ titles had non-technical terms, functional language, and focused on purpose rather than process. Emphasizing quantitative advantages of the technology (e.g., the technology resulted in lower pollution) also was advantageous. It seems feasible that a similar conclusion could be applied to graphs, in which the inclusion of specific information in titles would be beneficial to graph readers.

1.7.3 Captions

While titles are more ubiquitous regardless of the presentation source, captions are more commonly found in academic sources (textbooks, professional journals) than in other sources such as magazines (Kosslyn, 2006). Captions serve to increase the clarity of the graph as well as to direct attention or guide a reader’s processing of the graph (Kosslyn, 2006). Captions can be necessary when presented with ambiguous graphs, but Schield (2006) notes that captions in USA Today did not necessarily clarify whether the percentages given on the graph were parts of a subset or wholes. In general, though, in drawing a reader’s attention to what is important, the text in the caption connects the overall description of the graph with a statement relating to the conclusions (Cleveland, 1994; Kosslyn). (The good titles for technical papers served this function by using specific quantitative information (Senda et al., 2004).) However, while a caption can include a comprehensive description of “everything that is graphed” (Cleveland, p. 55), one of the risks is including too much information. However,

---

14 However, Kosslyn notes that the benefits of “clarifying and directing attention” must somehow be served whether or not a caption is present (p. 188).
Cleveland points out that this is not a common error with captions. It is also perhaps an error that is judged by the skill level of the graph reader and graph creator.

1.8 Summary

The research evidence shows that some degree of competency is present for certain types of graph interpretation tasks, specifically, table look-up and discrete comparison of values. However, there are still high error rates, especially as the difficulty level of the task increases. The difficulty level of the interpretation task is not independent from the other aspects of the representation. Specific graph characteristics (including appropriately applied graphing conventions) also influence performance. Overall, graph, task, and reader characteristics are not independent of one another. All of these factors are interacting within the representational format.
Real world events are influenced by many factors. It is most unlikely that phenomena in our world can be explained by just looking at the influence of a single variable. Instead, the effect of a particular independent variable (IV) often depends on a variety of other factors, such as what particular levels of other variables are present (Schaefer, 1976). For instance, the effect of high humidity on the heat index is greater at higher temperatures than at lower temperatures (see Figure 2.1). In statistics, such interrelated effects are conveyed by interactions among IVs. As “the world is not bivariate” (Wainer & Velleman, 2001, p. 316), considering multiple IVs and how the IVs interact is critical.

Figure 2.1. Graph depicting interactions of common IVs (humidity and temperature). Interactions are ubiquitous in our world. Here, the graph shows the interaction between humidity and temperature, in which humidity has a greater effect on the heat index at higher temperatures.
When at least two IVs are considered, it is possible to ask whether they interact, that is, whether the effect of one variable depends on the particular value of another variable. Correctly recognizing the impact of multiple causal variables is a critical part of scientific reasoning. Of course interactions are not limited to two variables, and variables are not limited to two levels. However, given the paucity of work on people’s understanding of interactions, it seemed wise to focus on the two IVs interactions case, especially since there is evidence that many people fail to understand even this simpler case (highlights will be reviewed in section 2.3). Surprisingly, graph-focused research pertaining to these areas is limited. Still, meaningful clues about people’s understanding of the concept of interactions come from other research areas such as statistical education and experimental design in inquiry learning.

2.2 Representing Data with Two Independent Variables

Figure 2.2 provides a schematic that exemplifies a possible relationship between two IVs as well as possible representational formats. In the presented example, the speed of widgets (the DV) depends upon two IVs: the environment and the material. This critical information can be represented by text (Figure 2.2, left panel). Another representation of the same information is the table presented in the middle panel. Finally, the same information about how the two IVs interact with the speed of widgets can be represented using a graph (Figure 2.2, right panel). The representation format does not indicate whether this relationship is causal or non-causal.
Widgets made of iron move quickly in a wet environment but slowly in a dry environment. Widgets made of plastic move slowly regardless of whether they are in a wet or dry environment.

**Figure 2.2.** Multiple representational formats of an interaction between two independent variables. The same information can be represented in text, tables, or graphs.

*Note:* A line graph format was chosen following Kosslyn’s (2006) recommendations in order to emphasize the pattern. The additional information readers know about the materials domain clarifies that IV_A is not continuous.

All three formats represent the same information, but as indicated in Chapter 1, the accessibility of an interpretation depends on the format and the interpreter’s skills. In the current research, the experiments present information in two of these formats: text-based statements and graphs. Each presentation format contains the information relating the IVs to one another and to the DV. The particular pattern demonstrated in the graph in **Figure 2.2** is just one of the possible relationship patterns between two IVs. For instance, **Figure 2.3** shows another possible pattern in which only the type of environment impacts the widget speed with widgets of both materials being faster in a wet environment.

### 2.2.1 Defining Interactions

The data patterns seen in **Figures 2.2** and **2.3** differ in terms of which IVs influence the DV. In **Figure 2.2**, IV_A (Environment) and IV_B (Material) both influence the
DV. In addition, the influence of $IV_A$ depends on the level of $IV_B$ —— this is what defines the concept of an interaction. In Figure 2.3, $IV_A$ (Environment) still affects the DV, but there is no effect of $IV_B$ (Material), and the influence of $IV_A$ is the same for both levels of $IV_B$. There is no interaction present in the data pattern represented in Figure 2.3.

**Figure 2.3.** Graphical representations of 2 IVs that do not interact.

*Note:* In contrast to Figure 2.2, a bar graph was chosen to represent the data pattern depicted here because there was no interaction between the IVs and the IVs were categorical.

There are multiple ways to talk about the concept of an interaction. **Box 2.1** presents Keppel and Wickens’ (2004) ways of describing statistical interactions.
An interaction is present when

- “the effects of one independent variable on behavior change at the different levels of the second independent variable”
- “the values of one or more contrasts in one independent variable change at the different levels of the other independent variable”
- “the simple effects of one independent variable are not the same at all levels of the other”
- “the main effect of an independent variable is not representative of the simple effects of that variable”
- “the differences among the cell means representing the effect of factor A at one level of factor B do not equal the corresponding differences at another level of factor B” [a student’s definition avoiding jargon]
- “the effects of one of the independent variables are conditionally related to the levels of the other independent variable”

Box 2.1. Alternate definitions of interactions (Keppel & Wickens, 2004). Although all conveying the same concept, the alternate definitions that Keppel and Wickens present for the concept of statistical interactions illustrate the nuances and the conceptual complexity of the concept.

Of interest is how they offer different ways to state the concept. It is known that conceptual understanding is more likely to occur if a concept is presented in many ways, and so the variations in the textual descriptions can help engender learning. However, it is clear that all of the definitions include the idea that it is necessary to consider both IVs simultaneously in order to understand their influence on the DV.

2.3 Working with Multiple Independent Variables and their Interactions

Although situations in which two (or more) IVs jointly influence a dependent variable are ubiquitous, authors typically comment on the difficulty of working with
interaction effects without providing empirical evidence to support this difficulty. There is little direct research on the topic of interactions, either as the concept presented in language or in graphs. Fortunately, there are clues in the cognitive strategies and statistics learning literatures. Educational researchers who assess students’ understanding of ANOVA concepts point out the persistent nature of students’ difficulties with these concepts (e.g., Wainer & Velleman, 2001). Some inquiry learning tasks, such as those involving experimental design and inductive generalization, present multiple IVs and thus consider data patterns that involve interacting variables.

2.3.1 Interaction Experiences in Statistics Courses

Since the concept of interactions is central to the idea of a factorial ANOVA, students encounter the concept of interactions in their statistics courses. They are taught that the comparison of group means is a way to evaluate and interpret the outcome of an experiment.\(^{15}\) The group means usually are presented in both numerical tables and graphs. It is therefore reasonable to expect that students encounter graphs with interacting variables in their statistics coursework.

Still, many undergraduates are not taught much about interactions in statistics courses. A survey of American undergraduate psychology programs (Friedrich, Buday, & Kerr, 2000) revealed that 40% of elite\(^ {16}\) and 60% of general higher education schools dedicated no more than one class period to factorial ANOVAs in the statistics course for

\(^{15}\) This will be considered in the discussion. Rosnow and Rosenthal (1989) argue against this approach.

\(^{16}\) Elite and general status of schools was determined by national rankings (Friedrich et al., 2000).
psychology students. Related tests that serve to explore the interactions received even less coverage (for post-hoc tests, 70% had no more than one class; for a priori contrasts, 90% spent no more than one class). Around 80% of the advanced statistics courses spent 2 weeks on factorial ANOVAs, but these advanced courses were not readily available. Only 41% of the departments in their sample offered advanced statistics, and only 6% of the departments required it. Thus, only a small subgroup of undergraduate students had prolonged class coverage of content material related to interacting variables.

A concern remains, however – what are students understanding and taking away from these courses? It appears that even when the concept of ANOVAs and interactions is a main emphasis of the course, students still have difficulties working with interacting variables. Green (2007) offers us her finding about her students’ learning ANOVA concepts, calculations, and interpretations. After intensive in-class and out-of-class work with data and interactions, Green’s students did well on specific, closed-response conceptual questions about interactions. During the course, students improved their formal write-ups of ANOVA results; they were able to adapt to a formulaic and highly structured approach to writing an APA style results section. However, at the same time, informal written and verbal descriptions of the results did not improve. It was still difficult for students to communicate the results of their ANOVA analyses. The undergraduates could not fluently and concisely summarize the relationships and interactions among variables despite the extensive course instruction on conceptual understanding of relationships among variables. Thus, independent of working with
graphs, conceptually describing the interaction remained difficult. (Using graphs to help communicate did not reduce this difficulty.)

### 2.3.2 Inductive Learning and Experimentation

Difficulties relating the interaction of variables to data patterns have been considered in studies where individuals are asked to combine inductive and inquiry learning strategies in order to design experiments or draw evidence-based conclusions. The assignment of the effects of causal variables in a multivariate system is one focus of inquiry learning research (e.g., Novick & Cheng, 2004; Kuhn, Black, Keselman, & Kaplan, 2000). Other scientific reasoning skills (e.g. hypothesis generating and testing) are necessary precursors to obtaining data and developing inferences about the roles of the variables. Systematically designing controlled experiments in this context allows participants to discover information about the relationships among the variables. However, participants do not always generate controlled experiments and thus do not avail themselves of the relevant data (Klahr, 1996; Klahr & Simon, 1999; Wilhelm & Beishuizen, 2003). One difficulty in doing so is that individuals may focus on achieving outcomes rather than evaluating variables (Kuhn et al.). This tendency emphasizes the necessity of focusing on the variables themselves, and not just the levels of the variable. Focusing only on the levels or only on achieving particular outcomes makes understanding the role of the variable more difficult.

In the above types of inquiry learning scenarios, participants are presented with variables (usually as either representations or simulations) and asked to determine how
the variables affect an outcome. Participants select a combination of the variables and levels, and then they are given numerical results reflecting what the DV would be if their experiment were conducted (e.g., Wilhelm & Beishuizen, 2003). This is in contrast to the current research, in which the data patterns and relationships are all given and do not need to be discovered. By limiting the kinds of processes required, there are fewer options for where errors could arise.

As an example of inquiry learning research, Wilhelm and Beishuizen (2003) compared the ability of undergraduates to uncover interactions among variables in different content domains. One of the conditions (as shown in Figure 2.4) involved familiar, concrete variables (bikes) that interacted in plausible ways.

The other condition used abstract variables (e.g., each IV was a different shape and the available levels were colors) for which domain knowledge would not have an impact on interpretations. Thus, the conditions varied based on which domain (familiar or abstract) was employed.

There were five IVs within each condition. Two of the IVs had no effect on the DV, one IV (with three levels) had a curvilinear effect, and the final two IVs (each with two levels) interacted. The interaction of these last two IVs is shown in Figure 2.4; this is the interaction pattern that individuals could plausibly discover during the inquiry learning task. Note that no graphs were presented in their research. The graph is used

---

17 The textual description given by Wilhelm and Beishuizen (2003) did not match the tabular presentation of what was supposedly the same data. The textual description is used here.
here to present one of the possible data patterns participants could conceivably discover.

Figure 2.4. Pattern between two familiar IVs (Type of Bike, Breakfast Location) that was discoverable in Wilhelm & Beishuizen’s (2003) self-directed inductive learning task. Overall, participants were unable to specify the existence of the interaction. Participants did not see this graph; rather, this was the data pattern that the interacting variables generated. The depicted interaction occurred at each level of a 3rd IV, which had 3 levels and showed a curvilinear effect on the DV.

Students were generally successful at testing for, detecting, and describing the curvilinear effect in both conditions (Wilhelm & Beishuizen, 2003). Picking up on this relationship corresponds to successfully comparing different values of a single IV in a graph. However, it was much harder for them to work with the interacting variables. Only about half of the undergraduates generated experimental designs that could provide evidence of the interaction. Further, they did not use the available evidence (as generated by the computer program that reported the results of their experimental design) in their inferences about the variables they manipulated. Only 6 of the 18
students working with the familiar IVs (e.g., the biking scenario) and 4 of the 26 of students working with the abstract IVs mentioned that the variables interacted, and only one student correctly described the interaction (Wilhelm & Beishuizen, 2003). Students in the familiar condition demonstrated greater understanding of the interaction as well as the absence of effects for some of the IVs than students in the abstract condition.

In a similar study, Schunn and Anderson (1999) presented content-rich experimental design scenarios to experts (faculty members in psychology) and novice undergraduates. Participants were asked to generate combinations of variables that could test how the IVs (e.g., spacing of practice, number of repetitions) affected the DV of word recall. The participants chose levels of each variable using a computer program. The program then simulated the running of the participant’s experimental design. Participants were asked to evaluate theoretical explanations based on the simulated data that the program reported. This required determining how IVs influenced the DV, including how the IVs interacted if applicable. Schunn and Anderson note that they specifically used tables to present the simulated data outcomes to the participants because they hypothesized the tables would be easier to understand than graphs.

Experts were generally successful at discovering main effects in this content-rich experimental design situation (Schunn & Anderson, 1999). Undergraduates were also fairly successful detecting and describing main effects as they inferred approximately 75% of the main effects demonstrated in the simulated data from their designed experiments. These “hits” for main effects are depicted in Figure 2.5. The authors note
that this success with interpreting main effects may overstate interpretive competence because the main effects were in agreement with an intuitive comparison of the variables. In contrast, the interactions were not supported by such intuitive relationships.

Overall, the interactions were much harder. Understanding interactions can be considered from two perspectives: correct detection and false alarms. Psychology faculty could use their general expertise with experimental designs to correctly detect interactions. The faculty who were studying the task domain (memory) had an additional advantage. Their domain expertise helped them recognize when an interaction was not reliable or meaningful (and thus avoid false alarms). The false alarm rate for interactions that is reported in Figure 2.5 comes from the faculty with only task expertise and not domain expertise. The rate of correctly detecting interactions was notably lower among undergraduates than among experts. Further, undergraduates’ rate of hits on interactions was not distinguishable from their rate of interaction false alarms.
Figure 2.5. Effects (main effects and interactions) reported by experts and undergraduates based on experimental data they generated (based on data from Schunn & Anderson, 1999). Their research further divided participants into multiple groups. Undergraduates with higher math SAT scores had higher rates of reported interactions (both hits and false alarms). Psychology faculty who studied memory (the content domain in which the tasks were conducted) had no false alarm reports of interactions.

The relevant data from Schunn and Anderson (1999) are presented in Figure 2.5; it shows that undergraduates were not able to differentiate actual interactions in the data from interactions that did not have large enough effects to be reliable or meaningful. Overall, even among those who generated appropriate experiments and thus were presented with relevant simulated data, undergraduates were half as likely to correctly interpret 2-way interactions as the faculty in psychology. In addition, the undergraduates’ interpretations about interactions that the undergraduates did make tended to be qualitatively different from the experts’ interpretation. Undergraduates were more likely to predict simple quantitative interactions than experts, who tended to rely upon qualitative versions of typical interaction patterns (Schunn & Anderson).
2.4 Which Data Patterns Have Been Investigated and/or Compared?

In each of the inquiry learning studies discussed earlier, the participants were tasked with determining the relationships among multiple independent variables. Participants had considerable difficulty determining the understanding the interactions among the IVs from the simulated data (Wilhelm & Beishuizen, 2003; Schunn & Anderson, 1999). In each inquiry learning study, however, one should note that only a single underlying data pattern was used. The data patterns that participants could discover with hands-on experimentation (albeit, often on a representational or simulated level) are the same data patterns that can be discovered when interpreting graphs.

The data pattern participants in Wilhelm and Beishuizen (2003) could discover was shown in Figure 2.4; the slopes of each level of $IV_A$ were of different magnitudes and opposing directions (the simple effects model presents this as data pattern 11 [$IV_A$ Complexity Level 5, $IV_B$ Complexity Level 4] in section 2.5.3). They describe this data pattern as a disordinal interaction and one in which $IV_B$ has a reverse effect depending on the level of $IV_A$. Participants in Schunn and Anderson’s (1999) task could discover that the slopes of the two levels of $IV_B$ were in the same direction but of different magnitudes (the simple effects model presents this as data pattern 5 [$IV_A$ Complexity Level 3, $IV_B$ Complexity Level 3] in section 2.5.3). They describe the pattern as one in

---

18 Kuhn (2007) notes that one of the close connections experimental design and scientific reasoning shares with causal inference is their underlying reliance on an ANOVA-type model. This model organizes the inclusion of multiple variables. However, despite these connections as well as those outlined in this text, the two research domains have tended to be separate (Kuhn).
which “the drop-off with increasing delay is faster at smaller spacings” – that is, a
pattern in which the slope of one level of IV (the level with smaller spacings) is more
extreme than the other level of IV (p. 348).\textsuperscript{19} This interaction occurs at both levels of
another IV in their design.

Similarly, researchers in the graphing literature tend to select only one particular
graph pattern when creating the graph stimuli, and then the investigator manipulates
other features of the graph representation or interpretation task. Even when more than
one pattern has been used, researchers still tend to look at the effects of manipulating
the representational format or the interpretation task. Shah et al. (1999) used three
arbitrary data patterns (albeit with more levels for some of the IVs or more IVs) and
focused on changes in format for each data pattern. They did so without systematically
varying or addressing the effects of specific data patterns on participants’ accuracy.
Their original data patterns came from social studies textbooks, and it appears that they
used several data patterns in order to introduce stimulus and content variation into the
task.

Others have systematically used the classification of data patterns to introduce
variation into the stimuli. For instance, Peebles and Ali (2009) selected data patterns to
represent each of the patterns outlined in the traditional model presented in Figure 2.6.
However, Peebles and Ali did not explicitly compare performance on the different
patterns.

\textsuperscript{19} In addition to this textual description of the data pattern to be discovered, Schunn
and Anderson (1999) also provided a table of data values. I note, however, that these
two representations do not match each other (as was also the case for Wilhelm and
Beishuizen’s (2003) paper). Interactions are hard to describe.
Carpenter and Shah (1998) offer an important exception to these trends with their comparison of three different features of data sets; tying their comparisons together is the theoretical approach that graph reader’s encodings are based on what they are able to visually chunk from a graph. Two of their comparisons of data patterns are particularly relevant to considering graphs with two IVs. The first set of graphs they contrast involves crossover and non-crossover data sets. The terminology of cross-over data sets is descriptive: the two lines representing the levels of IVB cross one another. This implies that the relation between IVB\textsubscript{1} and IVB\textsubscript{2} is different at each of the two levels of IV\textsubscript{A}. Non-cross-over data sets have opposing slopes for the two levels of IVB. (One can get from a cross-over perspective to a non-cross-over perspective by switching the locations of the IVs; the impact of this on performance was discussed in section 1.6.1.2.) Thus, Carpenter and Shah (1998) focused on the relation between the slopes of the levels of IV\textsubscript{B} in order to identify how data sets differ from one another. With the cross-over data sets, Carpenter and Shah (1998) suggest that participants encode the relations of the two lines similarly, without incorporating the relationship between IV\textsubscript{B} and the DV. On the other hand, the non-cross-over data sets would engender a more detailed encoding of the relationships that made greater use of both IVs. Their measure of the complexity of the graphs relied on the notion of ‘distinct visual functions’ of different slopes.

\footnote{One can also think about the distinction between cross-over and non-cross-over data sets in terms of whether the interaction is disordinal or ordinal (Howell, 2009). Ordinal interactions depict patterns in which the functions for IV\textsubscript{B} do not cross – one level of IV\textsubscript{B} is above the other for all values represented in the data. Disordinal interactions refer to the same concept as cross-over interactions; the impact of IV\textsubscript{B} is opposite at different levels of IV\textsubscript{A}.}
Another data pattern feature that Carpenter and Shah (1998) investigated was the presentation of patterns that both did and did not have a common value shared between the levels of IVB. In this situation, one IV does not have an effect at one of the levels of the other IV. In keeping with the focus on how graph readers would chunk the data sets, Carpenter and Shah emphasize the salient visual features of the two perspectives. In one version, one level of IVB has a slope of 0. When the locations of the IVs are exchanged, the two levels of IVB meet at one level ofIVA.

Hurts (2009) also presented subjects with graphs with two IVs, each with two levels. Participants in his studies were asked to evaluate different types of interpretation statements about the relationship among the levels of variables on the graph. The statements asked participants whether main effects were present, and if so, whether there was an interaction (in a specified direction). Detecting the interactions was no harder for participants than detecting the main effects. In fact, some of the findings suggested that detecting the main effects required more effort. The difficulty of different types of interpretation depended on other aspects of the task, however. When graphs had meaningful contexts, Hurts found that particular data patterns (showing higher degrees of parallelism) helped participants detect main effects, but not interactions.

Surprisingly, there have been few systematic comparisons across the different patterns of interaction and their representations. Instead, as covered above, the data pattern is either tightly controlled or treated as a nuisance variable. Given the variety of tasks employed and other variables influencing graph interpretation, it is not feasible to
compare results across these studies. The present studies, especially the Speeded Graph-Statement Evaluation Task in Chapter 7, address this gap by systematically varying graph patterns within the same task. To do this, it is necessary to have an account of the possible data patterns.

2.5 Categorizing Graph Patterns

There are a variety of ways to categorize graphic data patterns. The traditional classification is based on the outcome of statistical tests. Other approaches have been employed within specific investigations – these approaches tend to be variations and/or subsets of the traditional classification. The approach presented here (in section 2.5.3) is based on the author’s analysis of how graphs display simple effects. It embodies a more nuanced model of the complexity of data patterns in graphs and therefore the possible effects of dimensions that contribute to the complexity of graph patterns. Table 2.1 lists the various classification schemes that will be considered in this section.
### Table 2.1. Graph classification schemes.

<table>
<thead>
<tr>
<th>Label of Classification Scheme</th>
<th># of Data Patterns Considered</th>
<th>Figure Showing Data Patterns</th>
<th>Scheme Used By</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traditional</td>
<td>6 or 8</td>
<td>Figure 2.6</td>
<td>Peebles &amp; Ali (2009)</td>
</tr>
<tr>
<td>Basic caricatures of two-way interactions</td>
<td>3: additive, cross-over, and effect/no-effect interactions</td>
<td>Figure 2.6 (graphs E, G, &amp; H)</td>
<td>Schunn &amp; Anderson (1999)^A</td>
</tr>
</tbody>
</table>

*Note. ^A Schunn and Anderson (1999) limited their classification attempt to graphs in which an interaction is present. Thus, they can be treated as a subset of the traditional classification scheme.*

#### 2.5.1 Classification Based on the Main Effects and Interactions (Traditional Classification)

When textbook authors and statistics instructors cover the two-factor independent measures ANOVA concepts and computations, the focus is on patterns based on the presence or absence of main effects and interactions. These can be summarized into six (or sometimes eight) possibilities. These patterns (see Schaeffer, 1976) are shown in **Figure 2.6**.
**Figure 2.6.** Traditional classification of graphs with 2 IVs based on the number of statistical effects. Graphs are categorized based on the number of main effects present and whether or not there is an interaction between the IVs. Schaefer’s (1976) outline of this scheme also acknowledged the different patterns produced depending on which IV has an effect; this is represented by multiple patterns in two of the cells.

In this classification, the prototypical patterns are determined by whether or not there is an interaction, and then within those categories, by the number of IV showing a main effect (0, 1, or 2). This generates 6 cells but that can be increased to 8 possible data patterns if further delineating the cases where only one IV has a main effect.

When only IV_A has a main effect, the resulting patterns are C and G in **Figure 2.6.** When only IV_B has a main effect, the resulting patterns are B and F in **Figure 2.6.**

A cautionary note: the categories as well as the language and terminology used to label these patterns are not universal. Wilhelm and Beishuizen (2003) use the term
‘reverse’ to refer to graphs similar to Figure 2.6, pattern F. Their terminology matches with the data pattern: one IV has a reverse effect at one level of another IV, relative to its effect at another level of the other IV (e.g., the effect of IV_A is negative at IV_B1 but the effect of IV_A is positive at IV_B2). Peebles and Ali (2009) describe this same pattern as “crossed and converging” (p. 2939). Figure 2.6, Pattern G also has a variety of labels that focus on describing the pattern: “effect / no effect interactions (i.e., effect at one level, no effect at another level)” (Schunn & Anderson, 1999, p. 360) or “one horizontal line and one sloped line” (Peebles & Ali). Despite the differences in terminology, the goal in each case is a descriptive label that identifies a particular data pattern. One can also consider the features of the graphs that are emphasized in the various descriptions of the pattern. The emphasized features can indicate how the data patterns are being categorized. For instance, when Peebles & Ali focus on the visual appearance of the slopes, they are taking a different conceptual approach than Schunn and Anderson’s emphasis on the presence or absence of effects.

2.5.2 Variations on the Traditional Scheme

The classification scheme utilized by Peebles and Ali (2009) focuses on physical characteristics of the data pattern. They list the common patterns of “parallel, crossed and converging lines, one horizontal line and one sloped line, two lines sloping at different angles, etc.” (p. 2939), but also make it clear that their list is not exhaustive.

---

21 The descriptions do not uniquely point to a single graph in the traditional classification. For instance, graphs with parallel lines could refer to both patterns B and D in Figure 2.6.
They indicate, however, that the six graph patterns chosen and used in their experiment cover the common relationships encountered among graphs with two IVs and one DV. Thus, it is likely that their six patterns line up with those used in the traditional classification scheme.

Focusing only on graphs in which an interaction is present, Schunn and Anderson (1999) identify three general patterns: additive effects (likely Figure 2.6, pattern H) cross-over interactions (Figure 2.6, pattern E), and effect/no-effect interactions (Figure 2.6, pattern G). Schunn and Anderson (1999) suggest that expert researchers can use these “caricatures” of the data patterns (p. 360) to guide their experiment-space testing and to convey the expected pattern of the data without looking at exact magnitudes of data values. However, they then add an additional level of “a simple quantitative interaction (i.e., effects are always in the same direction, but the magnitude of the effect varies)” (p. 360).

Carpenter and Shah’s (1998) investigation of graph patterns (cross-over vs. non-cross-over and shared point vs. horizontal line) offer different classifications of the graph patterns in the traditional model. These two features result from switching the location of the two IVs; the different graph patterns used by Carpenter and Shah are discussed in Section 2.4. Their model suggests that the visual chunks that graph readers can create from the pattern influence what relations are encoded. When similar lines are chunked together, less detailed information is available for interpretations about the role of other variables that had been collapsed during the chunking process. A large part of what defines the visual chunks in their approach is the slope of the line and its
similarity to other slopes. This emphasis on a line’s slope features prominently in the model based on the consistency of the simple effects that is developed here.

2.5.3 Classification Based on the Consistency of Simple Effects

The classification scheme developed in this thesis allows a more nuanced consideration of the complexity of the relationship between the two variables. The approach I am developing here moves beyond the traditional approaches by adding in the conceptual issue of whether the simple effects of a variable are consistent in sign and/or direction as a defining mechanism for the categories. These factors influence the complexity of the visual data pattern in a systematic way. In this model, increasing complexity of the data pattern is identified as a decrease in the consistency of the simple effects. (Box 2.2 explains the concept of simple effects, and Figure 2.7 illustrates the presence of simple effects in a graph.)

As the consistency of the simple effects with each other and with the main effect changes, there are systematic changes in the visual characteristics of the data pattern. As discussed in Section 1.6.1.2, visual characteristics and features such as perceptual grouping influence graph interpretation. Thus, the model does not currently distinguish the relative contributions of the perceptual features from the contributions of the conceptual relationships among the variables. The focus, though, is on the conceptual similarities, as those are relevant to the types of interpretation statements subjects will be asked to make in the current research.
Simple Effects

Simple effects of an IV are the effect of one IV when the other IV is held constant. For instance, one of the simple effects in a graph with two two-level IVs is the effect of IV\textsubscript{A} at IV\textsubscript{B1}. This amounts to the difference (with a sign and a magnitude) between IV\textsubscript{A1}\textsubscript{-}IV\textsubscript{B1} and IV\textsubscript{A2}\textsubscript{-}IV\textsubscript{B1}. One could also find the simple effect of IV\textsubscript{A} at IV\textsubscript{B1}. Similarly, when analyzing the main effect of IV\textsubscript{B}, you can find the simple effect of IV\textsubscript{B} at either IV\textsubscript{A1} or IV\textsubscript{A2}. The simple effect of IV\textsubscript{B} at IV\textsubscript{A1} is again the signed magnitude of the difference between IV\textsubscript{A1}\textsubscript{-}IV\textsubscript{B1} and IV\textsubscript{A1}\textsubscript{-}IV\textsubscript{B2}. Note that in comparing two values from the data and finding the difference among them, one of the IVs is held constant, and the effect of the other IV is considered. Additionally, note that because the sign of the difference is considered, the order of subtraction must be consistent.

With 2 IVs, each with 2 levels, there are 4 simple effects (IV\textsubscript{A} at IV\textsubscript{B1}, IV\textsubscript{A} at IV\textsubscript{B2}, IV\textsubscript{B} at IV\textsubscript{A1}, and IV\textsubscript{B} at IV\textsubscript{A2}) in addition to the 2 main effects (IV\textsubscript{A} and IV\textsubscript{B}) and the interaction (IV\textsubscript{A} x IV\textsubscript{B}). Two of these simple effects are labeled in Figure 2.7. To understand and interpret significant interactions, researchers turn to the simple effects as components from which the interpretation can be built.\textsuperscript{A}

Box 2.2. Explanation of simple effects. The statistical concept of simple effects is critical to the development of the theoretical graph space based on the consistency of simple effects. In addition to defining the concept, this box references the notation used to refer to simple effects.

\textsuperscript{A} Rosnow and Rosenthal (1989a, 1989b) have written extensively as to why this is inappropriate. Section 7.5 addresses this issue. From the perspective of participants’ responses in the current studies, interpreting cell means is appropriate.
Figure 2.7. Simple effects depicted on a graph. The quantities referenced underneath the graph (e.g., IVA1_IVB1) illustrate how the levels of IVA and IVB are crossed.

2.5.3.1 Overview of the simple effects model

The consistency of the two simple effect values for a given IV will be characterized as a) the same as each other or b) different from each other. Special cases occur where at least one of the values of the two simple effects is zero\textsuperscript{22}. This characterization (same, different, involving 0) is applied to both the signs (i.e., directions) and the magnitudes (i.e., absolute values) of the simple effect. When the sign and magnitude information for each simple effect of an IV are combined with one another, this approach generates a theoretical graph space that models the contribution of each IV to the depicted graph pattern. Thus, this model is based on the data patterns

\textsuperscript{22} Simple effects involving 0 are treated separately for several reasons. Zero is a special case as it is an unsigned number. Additionally, a simple effect of zero indicates the absence of an effect (i.e., an effect with a value of 0), which is qualitatively different from effects that are present and either in agreement or disagreement with one another.
that develop when $I_{VA}$ and $I_{VB}$ have systematic changes in the consistency of their simple effects. The following text builds up this model, which is presented in full in Figure 2.11.

The basic assumption underlying the model is that the more similar the simple effects for a single IV are to one another and to the main effect for that IV, the less complex the data pattern is. This assumption can be supported on several grounds. If the pattern at each level of an IV is the same, the conclusions can be generalized across the levels. It is only when the patterns diverge from one another (when an interaction is present) that it becomes necessary to describe each unique simple effect. Secondly, among graphs which do show interactions, having some shared features (e.g., the signs but not the magnitudes of the simple effects are the same) again reduces the explanatory load compared to having no shared features. Together, these factors refer to the length of describing the graph pattern. Thus, in the model, the complexity of the graph pattern is assumed to be related to the descriptive length.

Thus, one of the features of the simple effects consistency model is to lay out the ways in which simple effects relate to one another. Although the categories are initially ranked on theoretical grounds, this is an empirical question that is tested in the research presented in Chapter 6. The ranking of the complexity levels of each data pattern is based on the agreement of the simple effects (in terms of both their sign and their magnitude). Neither criterion is assumed to be more influential on how one would interact with the resulting data pattern.
After first addressing how the levels are based on the consistency of the simple effects, I turn to another source of evidence that is supportive of the model. Visual inspection of the resulting data patterns offers supporting evidence for this initial conceptual ranking. The data patterns change in systematic and progressive ways. The following discussion, along with the figures and tables, presents alternative ways of discussing the development of this approach.

2.5.3.2 Levels of simple effects consistency

It is helpful to start by allowing only one IV to take on different values. Figure 2.8 sets up the complexity levels for a single IV; the first step examines the magnitudes (or absolute values) of the two simple effects for an IV. (Remember that each simple effect can be represented as a difference between two data points and thus has both a sign and a magnitude.)
Figure 2.8. Relationships between simple effect signs and magnitudes for one IV differentiate the complexity levels in developed model of graph classifications. This representational approach to the classification model emphasizes the idea of simple effect consistency. The decision points about the values of the simple effect signs or magnitudes are enclosed in boxes. The information that leads to the levels is summarized in Table 2.2.

Each decision point is asking about the sign and magnitude of the difference between two points. Thus, if you are looking at IV_A and its simple effects, you are interested in the simple effect of IV_A at IV_B1 and the simple effect of IV_A at IV_B2. This chart asks about the differences generated in each comparison and compares their magnitude and signs to one another.

The magnitudes can either be the same as or different from each other. The next decision point concerns the signs of the simple effects – they can be the same, different, or involve 0. The left side of the figure depicts the cases in which the simple effect magnitudes are the same. When the simple effects are both 0, that is Level 1. Level 2 occurs when the simple effects both have the same sign and the same magnitude. Graphed data patterns at Level 6 (note the temporary jump in levels) have simple effects with the same magnitudes but different signs. The right branch depicts cases in which the magnitudes of the two simple effects for that IV are different from each other. When the signs are the same, this is Level 3. Level 4 has different magnitudes for the two simple effects because one simple effect is unsigned (i.e., 0) and the other sign is non-zero. This is an intermediate level as for one level of an IV, the simple effect is similar to the main effect, whereas the other level simply has no effect as opposed to an opposite effect. Finally, Level 5 has graphs with different simple effect magnitudes and different simple effect signs.

In the explanation of the model, the focus is entirely on the lower-level simple effects, independent of the main effects. In fact, though, the nature of the main effects
can be derived based on the information about the consistency of the simple effects. For example, if simple effects are opposite in sign but equal in magnitude, they cancel each other out, and there is no main effect (Level 6).

Instead of using the flow chart and explicitly representing the evaluation of the sign and magnitudes of the simple effects, the characteristics of each level can be represented by themselves (see Table 2.2).

**Table 2.2.** Complexity levels based on the sign and magnitude of the simple effects.

<table>
<thead>
<tr>
<th>Complexity Level</th>
<th>Are Signs of Simple Effects the Same?</th>
<th>Are Magnitudes of Simple Effects the Same?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>(0, 0)</td>
<td>(0, 0)</td>
</tr>
<tr>
<td>2</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>3</td>
<td>Y</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>(0, non-zero)</td>
<td>(0, non-zero)</td>
</tr>
<tr>
<td>5</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>6</td>
<td>No</td>
<td>Y</td>
</tr>
</tbody>
</table>

*Note.* When there are additional constraints on the signs or magnitudes of the simple effects, the required values of the simple effects for that complexity level are listed in parentheses.
From looking at the table, one can think about the difference in complexity levels arising as the signs and magnitudes become increasingly more different from one another. In Table 2.2, Level 1 involves no effects (all effects: main, interaction, and simple are 0). Next, Level 2 has both the signs and the magnitudes of the simple effects the same. Level 3 has signs, but not magnitudes, the same. Levels 4 and 5 both represent the situation in which neither the signs nor the magnitudes are the same. Level 4 is presumed to be less complex than Level 5 because in Level 4, the difference arises because one of the values is 0, in contrast to Level 5 where the difference is signs opposing and there are two non-zero magnitudes.

Level 6 (with its magnitudes but not signs as the same) presents the alternative pairing to Level 3 (for which signs but not magnitudes were the same). The jump in complexity levels occurs in part because of the trend in the numerical and graphical patterns presented as one increases the complexity level under consideration. The decision was made to place Level 6 last for two reasons. First, when represented geometrically (as will be discussed next in reference to Figures 2.9 and 2.10), the graphical representations change in a systematic pattern that continues through the chosen order of the levels. Second, the ordering corresponds to the length of the description needed to convey the information from the data pattern.

Arguments can be made for other orderings. For instance, graphs at Level 6 depict a greater degree of symmetry than graphs at Level 5. The regularity of the

---

23 It is possible that the similarity of the signs of the simple effects are not empirically weighted the same as the similarity of the magnitudes of the simple effects. If that turns out to be the case, a re-ordering of the complexity levels would be required, but no re-generation of the different patterns would be necessary.
pattern could conceivably make Level 6 easier. However, the systematic, monotonic
nature of the geometric changes between the levels as well as the descriptive
complexity of the graph pattern resulted in the chosen ordering of the levels. The test
of the model in the research presented in Chapter 6 will empirically address these
questions.

Not until Level 5 did the signs of the simple effects become opposites of one
another. This becomes clearer when considering the graphical representation of the
patterns at each of these complexity levels.

2.5.3.3 Graphical representation of each level of simple effects consistency

In this discussion and explanation of the model thus far, no reference has been
made to how these levels appear graphically. This is because the graphical
representation of the level depends on which IV is being manipulated. Changes in the
simple effects of IV_A (the IV on the x-axis) are represented as changes in the slopes of
the IV_B levels. Figure 2.9 systematically changes the complexity level of the simple
effects of IV_A – as one progresses from left to right, the slope of IV_{B2} changes. This
results in changes to the simple effects of IV_A.

![Figure 2.9](image-url)

Figure 2.9. Graphical representation of changes in IV_A’s complexity level. Increasing
complexity of IV_A can be represented as changes in the slope of one level of IV_B. (The
depicted graphs at Levels 1 & 2 (with their coincident lines) are at complexity level 1 for
IVB. The depicted graphs at Levels 3 – 6 are at complexity level 3 for IVB.) The slope of IVB1 is held constant in each graph.

Starting with equal slopes at Levels 1 and 2, the slope of IVB2 changes in equal steps (step size = 2) until, at level 6, it is equal in magnitude, but opposite in direction to the slope of IVB1. This is the Level 6 data pattern in which the simple effect magnitudes are equal, but the directions are opposite. The intermediate graphs at Levels 3 – 5 are equidistant from one another and the endpoints.

*Note:* Gridlines and numerical units for the slope are added for easier reference; they do not represent actual slopes used. Graph stimuli were bar graphs without gridlines. IVB1 is a solid blue line, and IVB2 is a dashed red line in this representation. IVA1 is assumed to be at the left endpoints, with IVA2 at the right endpoints of each line segment.

Changes in the simple effects of IVB (the IV in the legend) are represented as changes in the relative spatial locations of the two levels of IVB; this amounts to changing where the two levels of IVB intersect with one another. Alternately, changes in the simple effect of IVB could be considered as vertical translation of at least one of the levels of IVB. Figure 2.10 systematically changes the complexity level of the simple effects of IVB – as one progresses from top to bottom in the figure, the relative location of IVB2 changes.

Table 2.3 presents information on the different representational instantiations of each level for IVA and IVB. The description considers the consistency of the sign and magnitude of simple effects at that level. Then, the remaining columns offer text-based descriptions of the changes illustrated in Figures 2.9 and 2.10. Reading these descriptions, it becomes clear that changing the complexity of the simple effects for IVA changes the slope of the lines and that changing the complexity of the simple effects for IVB changes the relative spatial locations of the lines.
Figure 2.10. Graphical representation of changes in IVB’s complexity level. Increasing complexity of IVB can be represented as changes in the relative spatial location of one level of IVB. In the constraints in place in the model, IVB1 has a constant slope. Thus, IVB2 systematically changes its relative spatial location going down a column in the theoretical graph space. IVB2 is placed relative to IVB1. Level 6 is predefined, as it is where there is no main effect of IVA – thus, IVB2 must cross at the midpoint of IVB1. The remaining graphs are produced by adjusting the values in equal step sizes (steps that are ¼ of half the midpoint of IVB1).

Note: Gridlines and numerical units for the slope are added for easier reference; they do not represent actual slopes used. Graph stimuli were bar graphs without gridlines. IVB1 is a solid blue line, and IVB2 is a dashed red line in this representation. IVA1 is assumed to be at the left endpoints, with IVA2 at the right endpoints of each line segment.
Table 2.3. Conceptual and perceptual description of each complexity level in the hypothesized graph space based on the consistency of simple effects.

<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
<th>What does this look like for IV_A?</th>
<th>What does this look like for IV_B?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>simple effects are the same as one another and as the main effect (all effects are 0)</td>
<td>Both lines have slope of 0</td>
<td>Coincident lines</td>
</tr>
<tr>
<td>2</td>
<td>simple effects are the same as one another and as the main effect (all effects non-zero)</td>
<td>Lines have same non-zero slope</td>
<td>Parallel lines</td>
</tr>
<tr>
<td>3</td>
<td>simple effects and main effect all have the same sign, but different magnitudes</td>
<td>Lines are not parallel, but slopes have same sign</td>
<td>Lines do not touch, but are not parallel</td>
</tr>
<tr>
<td>4</td>
<td>one simple effect same sign and larger magnitude than main effect, other simple effect is 0</td>
<td>One and only one line has a slope of 0.</td>
<td>Lines touch at an endpoint but do not cross</td>
</tr>
<tr>
<td>5</td>
<td>simple effects are opposing signs, but different magnitudes</td>
<td>Lines have slopes in opposing directions, but of different magnitudes.</td>
<td>Lines cross, but not at the midpoint</td>
</tr>
<tr>
<td>6</td>
<td>simple effects are opposing signs and same absolute value of magnitude, canceling out for a main effect of 0</td>
<td>Slopes are equal absolute magnitudes and opposing directions</td>
<td>Lines cross at their midpoint</td>
</tr>
</tbody>
</table>

In summary, while the simple effects initially replicate each other in levels 1 and 2, they then progress through the following levels: same direction but different magnitudes, one present and one absent, different direction and different magnitudes, and finally, different directions and the same magnitude.
2.5.3.4 The simple effects consistency model for both IVs

Having thus considered the simple effects of each IV independently, it is important to recognize that the interest is on data patterns with two IVs, each of which can be independently manipulated. Thus, the complexity levels for both $IV_A$ and $IV_B$ need to be considered simultaneously. When the graphical representation of the complexity levels was considered, Figure 2.9 held the complexity level of $IV_A$ constant in order to consider changes in $IV_B$’s complexity. Figure 2.10 held the complexity level of $IV_B$ constant in order to consider changes in $IV_A$’s complexity. Letting the simple effect complexity level of both IVs vary simultaneously leads to the full model. The full model has $IV_A$ patterns represented on one dimension (different columns) and $IV_B$ patterns represented on the other dimension of the table (different rows).

The entire graph space is presented in Figure 2.11; it is derived by crossing the complexity levels for $IV_A$ with the complexity levels for $IV_B$. For purposes of explanation and initial testing of this model, several constraints are imposed. In particular, $IV_{B1}$ always has a positive slope, and it is assigned to be constant from one graph to another. In addition, magnitudes at $IV_{B1}$ are required to be greater than magnitudes at $IV_{B2}$. Together with other assumptions (see Section 2.7), the number of possible patterns is reduced. However, changing these constraints is possible - similar graph spaces could be laid out under other choices for the constrained values. Within these constraints, the 20 graphs depicted in Figure 2.11 cover the entire space.

The theoretical space generated with this model can be split into two sections. The one section in the upper left of Figure 2.11, composed of complexity levels 1 and 2
for both IVs, does not involve interactions between the IVs. The other section, composed of complexity levels 3 – 6 for each variable, has interactions. 

For reference purposes, the graphs are numbered from left to right, top to bottom. The data patterns are presented as line graphs in Figure 2.11 as lines offer a more transparent display of the interactions (e.g., Kosslyn, 1993, 2006). (Figure 2.12 presents the data patterns as bar graphs, which are the stimuli used in the studies in Chapters 5 and 6.)

Thus, the theoretical predication is that increasing the complexity of the data pattern (as measured by the consistency of the simple effects) will have a negative effect on performance. Working with either the line graphs (Figure 2.11) or the bar graphs (Figure 2.12), as one moves along the levels (either the rows or columns of the table), the complexity of the relationship between the data points is theorized to increase. This is a theoretically derived space. The empirical research in this thesis will inform the psychological ordering of the complexity levels.

24 It is not logically possible to cross complexity levels 1 and 2 with levels 3 - 6. The dimensions and features required at the higher levels require interacting graphs, but the features at the lower levels exclude interactions. Once one introduces differences in the slope (as either direction and/or magnitude) to one level of an IV, it is impossible for the other IV to still have both of its simple effects and magnitudes be the same for each level. For instance, changing the slope of IVB₂ (such that it now differs from that of IVB₁) results in the two levels no longer being parallel. This indicates that the simple effects of IVB are no longer the same as one another. Thus, when the simple effects of IV₁ are not identical to one another, the simple effects of IV₂ cannot be identical to one another either. This results in the model being separated into the 2x2 region of graph patterns with no interactions (based on complexity levels 1 – 2) and the 4x4 region of graph patterns with interactions (based on complexity levels 3 – 6).
Figure 2.11.
Categorization scheme based on the consistency of simple effects (line graphs). These dimensions (see Tables 2.2 and 2.3) create a space of 20 graphs.

Note: IVB1 is a solid blue line, and IVB2 is a dashed red line in these graphs. IVA1 is assumed to be at the left endpoints, with IVA2 at the right endpoints of each line segment.
Figure 2.12. Categorization scheme based on the consistency of simple effects (bar graphs). Stimuli were presented as bar graphs, although the systematicity in the patterns is more apparent in the line graphs (see Figure 2.11).

The graph patterns are later referred to with numbers: they are numbered from left to right, top to bottom.

The DV and scale on the y-axes are left off for better legibility; IV$_{B1}$ is represented in blue, and IV$_{B2}$ represented in red.
2.6 Comparing Graph Classification Schemes

It is relevant to consider where the classification schemes overlap. Looking jointly at both the traditional classification scheme and the simple effects consistency model, the increase in the number of categories is apparent. The consistency of simple effects classification scheme divides the possible graph space more finely compared to other classification schemes. Stepping back for a moment, Table 2.4 indicates whether the IV has a main effect or is involved in an interaction for each of the complexity levels.

Table 2.4. Complexity levels based on the existence of main effects and interactions.

<table>
<thead>
<tr>
<th>Complexity Level</th>
<th>Main Effect Present?</th>
<th>Interaction Present?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>Y</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>4</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>5</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>6</td>
<td>No</td>
<td>Y</td>
</tr>
</tbody>
</table>

From this representation of the levels, one can see that there is a tendency for progressively more effects to be present. However, it is also clear that relying on the presence of main effects and interactions to define categories does not distinguish
among levels 3, 4, and 5 in the simple effects consistency model. On the other hand, the graphs that are used in the traditional classification scheme can all be represented in unique cells of the simple effects consistency model (see Figure 2.13).

**Figure 2.13.** Overlap among traditional and simple effects consistency classification schemes. The 8 graphs from the traditional scheme are systematically located within the simple effects consistency graph space. The graphs are represented in the prototypical location, but as the traditional model is defined by the number of main effects and interactions, the letters also refer to the remaining cells in the model.
2.7 Assumptions

With a DV on a continuous scale, there are an infinite number of possible values for each level of an IV, and thus an infinite number of possible data sets. This is the case even with only one IV. Thus, in order to classify possible patterns, some decisions need to be made about the point at which data patterns be grouped together. In setting up the classifications of data patterns, several assumptions or processes are at play in order to narrow down the possible graph patterns to the possibilities displayed in each of the classification schemes. The key issues are how the chosen graph pattern can be assumed to represent the infinite number of possible patterns and the breadth of a selected prototypical pattern.

The discussion of effects is based on the presence of visual differences in the graphed data. Therefore, when there is a visual difference between levels of an IV, there is assumed to be a significant difference. That is, interpretation statements about the influence of the IV on each other and on the DV refer to qualitative differences in the DV. These qualitative differences are assumed to be present if, and only if, there is a visually detectable difference. Thus, in Figure 2.14, there are differences between IV_{B1} and IV_{B2} at IV_{A2} in each graph, but not at IV_{A1}. Similarly, IV_{A} has an effect at both levels of IV_{B} in Pattern 9, but IV_{A} only has an effect at IV_{B1} in Pattern 10. These are visually detectable differences on the graphs.
The issue becomes how to categorize the 4 patterns in the middle. This issue is avoided in the current studies by only using the prototypical patterns. However, with actual data, the intermediate data patterns are common.

On the other hand, when actual data is considered, this is not necessarily true.

Actual data involves variability and measurement error. It is not clear without further information at which point between Pattern 9 and Pattern 10 in Figure 2.14 one would decide that there was no effect of IV_{B} on IV_{A}. A visually detectable difference in a graph does not determine whether an effect is present. In the classification schemes, there is no consideration of quantitative measures of reliability and effect size. The differences between the group means and the effect sizes are not quantitatively evaluated. This applies to all of the classification schemes when graph patterns are presented without the statistical information or background content knowledge needed to evaluate effect sizes. In the current studies, this is handled by only using the prototypical data patterns.
rather than the intermediate cases. In reality though, at some point a graph will change from having an effect to not having an effect – this is realized in the change from a graph being classified as one pattern or another.

However, the other issue addressed by assumptions also relates to the extent of unique data patterns represented by the prototypical graph pattern in the classification scheme. It is possible to think about this issue in terms of the content of the graph. While there are effects of variable placement on interpretation (see Section 1.6.1.2), exchanging variables still presents informationally equivalent graphs. One assumption at play is that the locations of the two IVs on the graph are interchangeable. This is not explicitly spelled out in all of the schemes, but is implicit in Schaeffer’s inclusion of two possible patterns in some of the cells in Figure 2.6. In those cells, the visual pattern changes when the locations of the IVs are exchanged. However, in the other cells in the traditional classification, a similar pattern form is produced when the IVs are exchanged. The second assumption is that the levels of the IVs are interchangeable. Exchanging the levels of IVB changes the mapping between the symbols and the referents. For IVA, this appears as a reversal in the slope of a line.

2.7.1 Geometric Patterns

These latter two assumptions can be formalized (and extended) by considering the effect of geometric transformations on a particular graph pattern. Starting from a particular graph (e.g., the boxed in graph on the left of Figure 2.15), it is possible to obtain a family of graph patterns (Figure 2.16) – all of which possess shapes that can be
considered qualitatively similar. This judgment is a theoretical assumption made in the models and not one that has been tested empirically. The data patterns within a family are being grouped with the prototypical data pattern; this limits the number of forms in the classification.

Figure 2.15. Geometric patterns applicable to creating a ‘family’ of graphs. The processes can be represented in terms of the geometric operation or in terms of the conceptual changes in the graph representation and its mapping to the real world referents.
Figure 2.16. A family of graphs, based on geometric relationships. The graphs in this family are related to the initial graph (in the upper left corner) by geometric transformations (as outlined in Figure 2.15). The smaller shaded graphs represent an extension of the family by the same processes. Unlabeled transitions arise from combining processes already implemented. In the categorization schemes, a single graph is used to represent the entire family.

The variations within each family are related to one another by geometric transformations. Within a family, graphs differ by reflection over a vertical midline (equivalent to switching \( IV_{A1} \) with \( IV_{A2} \)), reflection over the angle bisector (equivalent to switching the levels of \( IV_B \)), and reflection over a horizontal line (equivalent to reversing the slopes of \( IV_B \)). These processes, both independently and in combination, generate graphs belonging to a single family. The graphs within a family are assumed to be more similar than graphs from different families, although there are perceptual and conceptual differences in both cases.
Like the assumption of effects based on perceptual differences, the assumptions about exchanging the variables and their levels (i.e., geometric transformations) apply to all classification schemes. The stated assumptions already collapse a variety of graphs into each specific pattern. In addition, however, as is clear by the different number of data patterns specified in each of the classification schemes, the reach or size of the family of graphs that a particular data pattern stands for varies based on which classification scheme is used. Some classification schemes group more data patterns under the prototypical pattern.
CHAPTER 3

OVERVIEW OF EXPERIMENTAL TASKS

Graphs involving interactions between two independent variables are the focus across the experiments in the current research because of the level of difficulty demonstrated with each component, the prevalence of this type of information and communication, and the scarcity of research focusing on this common representational format (as reviewed in the earlier sections).

Bringing an emphasis on the data pattern to research about graphs allows a more nuanced understanding of how higher-level interpretation tasks are being performed. In the series of tasks reported here, the interpretation statements deal with the type of inferences one might reach when thinking about results of a research study. Each task involves statements about the main effects of the IVs, the simple effects of IVs, and the interaction between the two IVs. All three of these types of interpretation statements require the graph reader to perform more than a single point-reading operation with the graph.

These three types of statements are presented in conjunction with different graph patterns. The data patterns depicted in these tasks are selected based on the simple effects consistency model (section 2.5.3). As it was set up, this model makes specific predictions about the influence of the data pattern on performance. Across the three studies, the influence of the different data patterns on graphical literacy is of key interest.
As a reminder, two aims were introduced in Chapter 1. Collectively, the following research serves to: 1) assess individuals’ ability to construct graphs with two IVs while simultaneously also addressing the impact of graph-relevant representational structure to graph construction and 2) evaluate the understanding and interpretation of graphical data patterns showing interactions. This latter aim can be more broadly considered in the context of the two studies – the first study examines the effect of exposure and learning opportunities (a statistics course) on performance, and the second study investigates the interaction between the graph pattern and the interpretation statement.

In addressing these goals, two representational formats for the relationships between two IVs are employed: graphs and written statements that interpret a graph’s content. Moving between these representational formats requires graphical literacy. The experiments address graph interpretation from complementary directions: generating a graph from a given set of interpretation statements (Figure 3.1, a) and evaluating a given interpretation based on a given graph (Figure 3.1, b and c). Both of these directions require participants to implement graph interpretation (or construction) skills.
3.1 Participants

All participants were undergraduate psychology students at a large research university known for its ethnically and socioeconomically diverse student body. The number of participants in each study is given in the respective chapters, along with
further details about participant backgrounds (e.g., math skill levels\textsuperscript{25}, age, etc.).

Participants were recruited either through their participation in the subject pool as a requirement for introductory psychology or through extra credit / movie ticket offers within classes. Numerical literacy was assessed for those recruited through the subject pool, using Obrecht, Chapman, and Gelman’s (2007) multiple choice modification of the numeracy scale form Lipkus, Samsa, and Rimer (2001). All participants signed informed consent forms that indicated that the research had been approved by the human subjects committee.

3.2 Tasks

In Chapter 4 (Draw the Data: Adding Graph-Relevant Information), the upper pathway in Figure 3.1 is pursued: participants generated graphs based on written descriptions of data. Initially, this evaluates their generative understanding of graphs. Within this study, the repeated construction of graphs by participants investigates which aspects of the task cause difficulty. A between-subjects manipulation on the 2\textsuperscript{nd} trial adds graph-relevant representational information to the task. The manipulation identifies the variables and locates them spatially on the graph axes. This manipulation was chosen to assess how the statement constraint satisfaction components of the construction task (generating data values that are in concordance with all of the statements) and representational components (e.g., the identification and placement of

\textsuperscript{25} Math level is considered for two reasons: 1) the interpretation statements make qualitative comparisons about quantities and 2) math courses are one context in which students encounter and learn about graphs. The math levels students report could reflect both interest and achievement.
variables – a part of graphical literacy) causally influence graph construction. Thus, the
experiments explicitly addresses ways in which graphical literacy can be increased.

In Chapter 5 (Effect of Statistics Education on Graph Interpretation) and
Chapter 6 (Speeded Evaluation of Graph-Statement Matches), participants’
understanding of graphs depicting two IVs and one DV is evaluated using true/false
judgments and multiple choice questions. In both of these tasks (the lower pathway of
Figure 3.1, b and c), participants are given a statement and given a graph. In addition,
Chapter 5 assessed interpretation abilities at the beginning and end of an introductory
course in statistics. Students had spent the few weeks immediately before the 2nd
testing point learning about ANOVA and interactions among variables.

Chapter 6 focused on the interaction between the data pattern depicted on a
graph and the type of interpretation statement. The task presented 400 trials in which
the 20 graph patterns used as stimuli were crossed with the 10 of the statements under
consideration. Delving into how these components interact is informative on its own
about students’ abilities with these topics, as well as allowing an understanding of
where difficulties might arise.
4.1 Overview

Rarely are participants asked to construct graphs based on qualitative descriptions of data patterns. Some of the required skills are those used when constructing a graph from a table of numbers, such as the using axes and representing data using spatial features. These are relevant skills whether starting with numerical data values or not. Other skills are unique to this task, such as determining how qualitative comparisons translate into data patterns. When working with the qualitative description of the data pattern, the emphasis is primarily on the relationship(s) between the variables.

Mevarech and Kramarsky (1997) gave eighth graders simple summaries about data patterns and asked them to construct graphs. Each of their scenarios involved one IV (study time) and one DV (grade). The four scenarios given to participants each involved a different functional relationship between study time and grades. For example, one of the summaries (shown in Box 4.1, Prompt A) described an increasing relationship between the amount of studying and an individual’s grade. Another scenario (Box 4.1, Prompt B) describes a graph in which there is an initial increase in grades as studying increased, and then a decrease in grades as studying continued. The other two scenarios described a decrease in grades when studying increased or no effect of studying on grades. While the scenarios did not demand linear functions or
parts, students tended to make an assumption of linearity in their constructed representations.

Scenario A:

“Sarah claimed that the more she studies, the better her grades are. Please construct a graph that represents Sarah's claim.”

Scenario B:

“Rachel, however, said that when she studies up to three hours, the longer she studies the better her grades; but, beyond three hours, she becomes tired and her grades are lower. Please construct a graph that represents Rachel’s claim.”

Box 4.1. Sample prompts that were given to 8th grade students for graph construction (from Mevarech & Kramarsky, 1997). Two of the four scenarios are presented in the box. The first scenario describes a data pattern in which increases in the IV (study time) correspond to increases in the DV (grades). The second scenario prompt describes a data pattern in which increases in the IV correspond first to an increase and then to a decrease in the DV.

In between the two testing times, students were taught the standard 3-week instructional unit about graphs. The instructional unit covered the idea of a number line, Cartesian graphs (including ordered pairs), and skills in constructing and interpreting graphs of various formats. These skills included the concepts of variables (DV & IV) along with different functional relationships between the variables. The graphs included both discrete and continuous data. However, there was not an emphasis on algebraic forms of the graphs, nor were students’ errors or misconceptions targeted during the instructional unit. Students improved after the instructional unit. Only a quarter of students represented the 4 graphs correctly at pre-test, but half of the students represented all 4 graphs correctly at the post-test.

The students’ incorrect graphs were characterized by the type of conceptual error that would have led to the production of that type of error. Mevarech and
Kramarsky (1997) refer to these errors as alternate conceptions of the concept of a graph. The most frequent error type was using a series of graphs. In this alternate conception of graphs, the 8th graders did recognize that both variables (i.e., time studying and grades) were important, but they had difficulty with the simultaneous combination of the IV and DV. Thus, for the scenario describing an increasing functional relationship, the students produced one graph that representing an increase in time studying and another graph that represented an increase in grades. In this type of error, the two variables were not linked to one another in the graphical representation, and thus they were not treated as independent and dependent variables with a functional relationship.

Mevarech and Kramarsky (1997) commented that students with higher levels of graphical literacy were less likely to include units on their axes. They suggest that this is because the students recognized the generality of the relationship described by the data pattern. This is a feature of graphing qualitative descriptions of a data pattern. Similarly, Schunn and Anderson (1999) noted that psychology faculty tended to use qualitative versions of typical data patterns when predicting results in experimental psychology research. These comments support the validity of using constructed representations as a means of distinguishing and characterizing different levels of graphical literacy.

Pilot testing for the current experiment using Mevarech and Kramarsky's (1997) scenarios with one IV and one DV in an undergraduate population revealed that students were correctly able to work with these functional relationships. However,
graphs with 2 IVs present additional challenges – including the requirement of relating two IVs to one another as well as to the DV. In addition, as more variables are involved and the data pattern becomes more complex, there are more statements used to describe the data pattern. Thus, in the current study, a similar task at a higher level of difficulty was used with undergraduates.

4.2 Current Study

In this study, undergraduates read a series of facts about the relationships among variables, and they were asked to construct a graph based on each series of facts. For example, Figure 4.1 presents one version of the first trial. Participants initially read several short paragraphs about the study that generated the data they would be graphing. Bulleted statements about that data were presented on the lower left, and blank axes were provided on the right for the participant’s construction of the graph. These aspects of the stimuli (the background context, the comparative statements, and the axes) will be covered in the materials section (4.3). On each trial, participants were to base their construction of a graph on the combined information from the series of eight statements. Each statement described one of the qualitative relationships among two two-level independent variables and a dependent variable. With an emphasis on the relationships among the variables, pre-specified units or values are not presented in the task.
Brief Summary of Instructions: Construct a graph based on the following statements.

Background:
In this study, researchers wanted to find out how perceived danger and bystander presence affected the frequency of helping. Subjects witnessed a staged argument that resulted in injury to one of the people before they had the opportunity to help.

For perceived danger, subjects saw an argument in which the type of injury indicated a) low danger (accidental) or b) high danger (intentional). For bystander presence, subjects were a) alone or b) accompanied.

Then, researchers measured the frequency of helping.

Statements:  

- On average, subjects who were alone helped more often than subjects who were accompanied.
- When viewing a low danger situation, subjects who were alone helped more often than subjects who were accompanied.
- When viewing a high danger situation, subjects who were alone helped more often than subjects who were accompanied.
- The effect of bystander presence when viewing a low danger situation was smaller than the effect of bystander presence when viewing a high danger situation.
- On average, subjects who viewed a low danger situation helped less often than subjects who viewed a high danger situation.
- When viewing alone, subjects who viewed a low danger situation helped less often than subjects who viewed a high danger situation.
- When viewing accompanied, subjects who viewed a low danger situation helped less often than subjects who viewed a high danger situation.
- The effect of perceived danger when viewing alone was larger than the effect of perceived danger when viewing accompanied.

Figure 4.1. Example of trial in graph construction task. This was one of the versions participants saw on the 1st of the 3 graphs they constructed. The full instructions are presented in Box 4.3. The 2nd and 3rd graphs referred to different context domains, and the other versions of each trial referred to different data patterns. On each trial, the 8 bulleted statements constrained the correct data pattern. Blank axes, on which participants were to construct their graph, were presented to the right of the bulleted statements.
This is not an easy task. It involves both constraint satisfaction as well as graph representation skills. In order to separate out these two components of the task, a between-subjects manipulation was used in a pre-test, intervention, post-test design. The manipulation consisted of adding graph-relevant information to the axes: the labels and levels for the IV and DV were placed on the graph. This reduces the demands posed by the graph representation component of the task. It was hypothesized that subjects would be more likely to construct typical graphs (i.e., graphs that represented the relationships among the variables) during the manipulation trial and the post-test in the Added Labels condition.

However, one can represent a value without representing it accurately. As this is a novel task, it was unclear how accurately participants would integrate the statements to complete the statement constraint satisfaction part of the task. Their satisfaction of these constraints was the process that resulted in the magnitudes of the quantities they graphed. Thus, accuracy of the relationships can be directly related to ability to satisfy the constraints by using the information from the statements. Thus no predictions were made about accuracy levels in representing the statements, but accuracy was not expected to be impacted by the Added Labels manipulation as that was targeted towards the representational components of the task.

As Mevarech and Kramarsky (1997) had found that the particular data pattern to be represented affected representational accuracy, various graph patterns arising when 2 IVs are present were included in the current research to reduce the pattern-specific effects.
4.3 Materials

4.3.1 Context Domains

Each trial started with background paragraphs that introduced the source of the data facts presented in the statements. These contexts were based on content that would be typical of introductory psychology textbooks. The background paragraphs described the two IVs and their levels as well as the DV. These variables were laid out in the context of an experiment that other researchers had run. The subsequent statements about the relationships among the variables are said to be the results of the experiment laid out in the introductory paragraph.

Each of the three trials used a different context domain. Box 4.2 gives the full introductory text of each context. All participants saw all three contexts in the same order (helping, emotions, and then hand grip). While the content and particular relationships changed for each graph, the statements and background paragraphs were structurally identical for each problem. As the introductory paragraphs were on the same page as the comparative statements and graph, participants could refer back to the introductory paragraphs to remind themselves about the variables and how they were part of the context.
Box 4.2. Introductory text read by participants for each of their 3 trials. The context domains were always presented in the same order. The introductory paragraphs for each of the context domains identified the variables that would be referenced in the bulleted statements. For reference to the variables within the text, IV_A was the first IV mentioned, and IV_B was the second variable.
4.3.2 Qualitative Comparison Statements

As seen in the example presented earlier in Figure 4.1, participants read eight bulleted statements about the data pattern on each trial. These statements qualitatively described how the levels of the DV were based on the different combinations of the independent variables: IV_A and IV_B. The statements were presented in an ordered fashion, although this ordering was not explicitly made apparent to the participants. As can be seen in Table 4.1, one independent variable (IV_B)\textsuperscript{26} was the focus of the first four bullet points, and the statements about the other independent variable (IV_A) followed. For each IV, the first statement indicated the direction of the main effect, the next two statements indicated the direction of each of the two simple effects, and the final statement described the interaction of the IVs. This last type of statement about the interaction focused on the role of the IV emphasized in that half of the statements; it was a comparison of the relative magnitudes of the two simple effects. Not every piece of information in the statements was unique; there was some redundancy.

---

\textsuperscript{26} IV_B refers to the second variable presented in the background paragraphs. Despite its presence in the first half of the statements, it is still referred to as IV_B because of the placement of the variables in the experimental condition (IV_B is represented using the legend) and consistency with the following chapters.
Table 4.1. Statements presented to participants about the data pattern which they were to graph. Statements were ordered by the IV they focused on, but participants only saw the text versions.

<table>
<thead>
<tr>
<th>Bulleted Stmt #</th>
<th>Sample Statement</th>
<th>Description of Statement Content</th>
<th>Syntactic Form of Statement</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>On average, subjects who were alone helped the same amount as subjects who were accompanied.</td>
<td>Main Effect of IV&lt;sub&gt;B&lt;/sub&gt;</td>
<td>IV&lt;sub&gt;B1&lt;/sub&gt; ? IV&lt;sub&gt;B2&lt;/sub&gt;</td>
</tr>
<tr>
<td>2</td>
<td>When viewing a low danger situation, subjects who were alone helped less often than subjects who were accompanied.</td>
<td>Simple Effect of IV&lt;sub&gt;B&lt;/sub&gt; at IV&lt;sub&gt;A1&lt;/sub&gt;</td>
<td>For IV&lt;sub&gt;A1&lt;/sub&gt;, IV&lt;sub&gt;B1&lt;/sub&gt; ? IV&lt;sub&gt;B2&lt;/sub&gt;</td>
</tr>
<tr>
<td>3</td>
<td>When viewing a high danger situation, subjects who were alone helped more often than subjects who were accompanied.</td>
<td>Simple Effect of IV&lt;sub&gt;B&lt;/sub&gt; at IV&lt;sub&gt;A2&lt;/sub&gt; for IV&lt;sub&gt;A2&lt;/sub&gt;, IV&lt;sub&gt;B1&lt;/sub&gt; ? IV&lt;sub&gt;B2&lt;/sub&gt;</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>The effect of bystander presence when viewing a low danger situation was smaller than the effect of bystander presence when viewing a high danger situation.</td>
<td>Comparison of Simple Effects of IV&lt;sub&gt;B&lt;/sub&gt;</td>
<td>IV&lt;sub&gt;B&lt;/sub&gt; had a ? effect for IV&lt;sub&gt;A1&lt;/sub&gt; than IV&lt;sub&gt;A2&lt;/sub&gt;</td>
</tr>
<tr>
<td>5</td>
<td>On average, subjects who viewed a low danger situation helped less often than subjects who viewed a high danger situation.</td>
<td>Main Effect of IV&lt;sub&gt;A&lt;/sub&gt;</td>
<td>IV&lt;sub&gt;A1&lt;/sub&gt; ? IV&lt;sub&gt;A2&lt;/sub&gt;</td>
</tr>
<tr>
<td>6</td>
<td>When viewing alone, subjects who viewed a low danger situation helped less often than subjects who viewed a high danger situation.</td>
<td>Simple Effect of IV&lt;sub&gt;A&lt;/sub&gt; at IV&lt;sub&gt;B2&lt;/sub&gt;</td>
<td>for IV&lt;sub&gt;B1&lt;/sub&gt;, IV&lt;sub&gt;A1&lt;/sub&gt; ? IV&lt;sub&gt;A2&lt;/sub&gt;</td>
</tr>
<tr>
<td>7</td>
<td>When viewing accompanied, subjects who viewed a low danger situation helped the same amount as subjects who viewed a high danger situation.</td>
<td>Simple Effect of IV&lt;sub&gt;A&lt;/sub&gt; at IV&lt;sub&gt;B2&lt;/sub&gt; for IV&lt;sub&gt;A1&lt;/sub&gt;, IV&lt;sub&gt;A1&lt;/sub&gt; ? IV&lt;sub&gt;A2&lt;/sub&gt;</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>The effect of perceived danger when viewing alone was larger than the effect of perceived danger when viewing accompanied.</td>
<td>Comparison of Simple Effects of IV&lt;sub&gt;A&lt;/sub&gt;</td>
<td>IV&lt;sub&gt;A&lt;/sub&gt; had a ? effect for IV&lt;sub&gt;B1&lt;/sub&gt; than IV&lt;sub&gt;B2&lt;/sub&gt;</td>
</tr>
</tbody>
</table>
Thus, in total, there were 2 statements about main effects (one for each variable, e.g., “On average, subjects who were alone helped the same amount as subjects who were accompanied.”). There were 4 statements dealing with all of the possible simple effects (e.g., “When viewing a high danger situation, subjects who were alone helped more often than subjects who were accompanied.”). In addition, there were then 2 interaction statements (one focused on the effect of each IV, e.g., “The effect of bystander presence when viewing a low danger situation was smaller than the effect of bystander presence when viewing a high danger situation.”). All the statements, in the order in which participants read them, are presented in Table 4.1.

The particular data pattern specified by the chosen graph (see Section 4.3.3) determined the direction of the relationships specified by the statements. Together, the eight qualitative comparison statements presented would completely constrain the data pattern.

A further note is necessary in regards to the interaction statements. A different form of these statements was unintentionally used for the pre- and post-test graphs than for the 2nd graph on which the manipulation occurred. Both forms27 of the

---

27 On the 1st and 3rd graphs, the interaction statements compared the effect of one IV at each level of the other IV. For example, “The effect of bystander presence when viewing a low danger situation was smaller than the effect of bystander presence when viewing a high danger situation.” On the 2nd graph, the statements referred only to one IV: job type, with its two levels of working alone or working in a group. For example, “The effect of job type when working alone was smaller than the effect of job type when working in a group.” Following up with participants after they completed the task, indicated that they did not notice anything unusual about the interaction statements on the 2nd graph, and that they interpreted it as if it was in the form presented on the 1st and 3rd graphs.
interaction statement could lead to the same interpretation. However, there were multiple valid interpretations or representations of the information in the statement. The main issue lies in whether or not a “bigger effect” is considered to refer to absolute magnitude or both magnitude and direction. This, along with anecdotal reports from students that they ignored these statements, lead to the decision to remove these statements from further consideration. Accuracy analyses will look only at the main and simple effects.

The statements did not constrain the absolute magnitudes of the quantities. The comparisons in the statements only constrained the ordinal relationships between the data values at the different combinations of the independent variables. Thus, participants could choose differing effect sizes when they constructed their graphs. In addition, participants were free to locate either variable along the x-axis and to represent the levels in any order. As will be seen later, the Added Labels condition presented an exception to this. Thus, visually different representations of the same data pattern were possible depending on how the participant assigned the variables and their levels to spatial locations on the graph.

4.3.3 Graph Data Patterns

The graph patterns used in this experiment were selected from the theoretical graph space outlined in Chapter 2. Particular patterns were chosen so that different areas of the graph complexity space would be present. As can be seen in Figure 4.2, two sets of data patterns were used. The first set used graph patterns 5, 12, and 18; the
second set used graph patterns 10, 13, and 20. A limited selection of possible patterns was used in order to be able to control the effect of graph pattern on performance.

**Figure 4.2** Graph patterns described by the sets of statements presented to participants. The 8 statements (see Table 4.1) presented on each trial were based off the relationships prescribed by the relevant data pattern. Thus, the information from the statements constrained which pattern was correct.

**Table 4.2.** Distribution of graph patterns across subject groups. There were 4 different lists of stimuli. Sets A and B differ in the particular graph patterns used. Within each set, the graph pattern used in the 1st and 3rd trials was counterbalanced, leading to original and reversed orders of the set. Each subject received only one list of stimuli. Analyses collapse across stimuli lists.
Table 4.2 summarizes how the graph patterns were counterbalanced. Patterns 12 and 13 always appeared on the 2nd graph, and the order of the 1st and 3rd graphs (Set A: 5 and 18 or Set B: 10 and 20) was counterbalanced. Thus, the variation in the orders counterbalanced the complexity of the data pattern across participants and conditions. Within each list, there was a version for the Added Labels condition and a version for the Control condition. The varying patterns were used to reduce the effect of particular graph patterns, thus the analyses will be collapsed across graph data patterns in order to examine the effect of the Added Labels manipulation on performance. Although limited sample sizes for the subgroups preclude the inclusion of graph pattern in the full set of analyses, I return to the issue of whether graph pattern impacted performance on this task in section 4.10.

4.4 Experiment Design

Each participant completed 3 graphs on a set of axes that was provided for them. Axes were presented to the right of the list of statements; all axes were 10 cm x 10 cm. The overall design of the task is represented in Figure 4.3. In the Control condition, the axes were blank for all three of the to-be constructed graphs. In the Added Labels experimental condition, the axes were pre-labeled for the participants on the second graph.
Figure 4.3. Experimental design for Draw the Data task. All participants received blank axes on which to construct their graph representation on the 1st and 3rd graphs. A between subjects manipulation presented half of the participants with added graph-relevant information on the 2nd graph in the Added Labels condition.
Figure 4.4. Added Labels manipulation. The 2\textsuperscript{nd} graph in the Added Labels condition provided additional graph-relevant structure: labeled variables and their levels were spatially placed on the graph. Image is rescaled for font legibility.

The level of representational support was manipulated between-subjects. As laid out in Figure 4.3, all subjects received the blank axes on the 1\textsuperscript{st} and 3\textsuperscript{rd} graphs as well as on the 2\textsuperscript{nd} graph in the Control condition. This level of graph representation consisted of a blank, unlabeled set of axes. In the Added Labels condition, graph-relevant representational information added to the graph. The variables and their levels were identified and spatially located in the legend, along the x-axis, and on the y-axis (see Figure 4.4). Thus, in this case, participants were constrained as to which variable (IV\textsubscript{A}) was represented on the x-axis.
4.5 Procedure

Packets were distributed to the participants. After completing the consent form, participants encountered the 1st graph (pre-test), the background questionnaire, the 2nd graph (with added labels or not), and then the 3rd graph (post-test). Upon starting the construction task, participants were instructed to complete the packet without looking ahead or returning to previous pages. The background information sheet was intentionally placed in between the first and second graphs so that participants could not see the added labels presented in the experimental condition.

For each of the three graphs, the instructions, background paragraphs, comparative statements, and axes were all presented on one page, as shown in Figure 4.1. The instructions were repeated at the start of each graph page. As shown in Box 4.3, the instructions emphasized that the statements were to be the source for the information participants would graph. Participants used the comparative relationships specified by the statements to construct their graph of the data.

**Instructions:**

You will be given information and asked to make a graph based on that information.

The background paragraphs tell you about a research study that collected data. Then, the statements describe the data values that were obtained by the researchers.

Based on the statements, your task is to make a graph that represents all of the information you were given in the statements. (Using numbers on the vertical axis is not necessary, but please be precise with your markings.) Some of the statements may have redundant or repetitive information. Combine the information from all of the statements and only the statements to draw your graph.

**Box 4.3.** Instructions were printed on each trial of the graphing construction task.
4.6 Participants

Forty-five undergraduate students (20 female; 25 male) participated in exchange for research credits in their introductory psychology course. All participants signed informed consent; the research was approved by the human subjects IRB. Participants attended a socioeconomically and ethnically diverse large university in the northeast. The majority (80%) were in their 1st or 2nd undergraduate year. One additional participant was excluded because the first trial was incomplete, and the remaining two trials were blank.

4.6.1 Math Levels

As the task requires the comparison of relational qualities as well as the construction of graphs, math background was taken into consideration. Half of the participants had taken a statistics course, possibly in high school. Pre-Calculus (or less) was the highest level of math reached by 38% of the subject. The majority of the participants had taken (or were currently taking) Calculus I or higher. Together with math SAT score (less than 600, greater than or equal to 600), numeracy, and if they were a math/science major, the math background variables were averaged to form a composite measure.

Participants assigned to the two conditions did not have significantly different \((t(43) = -0.90, p = .37)\) levels on the composite math variable or on the individual background variables (numeracy, math/science major, math courses, math SAT, \(ps > .18\)), each of which was coded as low or high. However, as seen in Figure 4.5, 30% of
the control group (centered math composite $M = -0.05, SE = 0.07$) had math levels above average, while 50% of the experimental group ($M = 0.04, SE = 0.07$) had math levels above average.

![Figure 4.5](image)

**Figure 4.5.** Math levels in the Control and Added Labels conditions. Means for each group were not significantly different, but there was a larger group of participants with higher composite math scores in the Added Labels condition. These math groups were collapsed into three levels: below average, average, and above average.

The impact of math (either in terms of highest math experience or the composite math background variable) will be discussed where applicable in each of the following sections.

### 4.7 Coding

In coding participants’ constructed graphs, both quantitative and qualitative categories were used. Each of these is considered in separate results sections. The qualitative components precede the quantitative components as they provide the context in which the quantitative aspects of the representation can be considered.
4.7.1 Evidence of Effort

Each problem was also assessed for evidence that indicated a participants’ response strategy and effort. Four areas were defined: 1) evidence of revision of the bar heights (crossing out or erasures), 2) revision of labels, 3) reading marks such as underlining in the background paragraphs or the comparative statements and 4) other writing, such as abbreviations and inequality statements.

The coding process resulted in an accurate/inaccurate coding for each of the 6 main and simple effect statements. When relationships were not represented, they were distinguished in terms of whether any of the other relationships were represented or not. Graphs which represented some – but not all – of these relationships were coded based on which relationships were present, and the missing relationships were coded as missing. Graphs that represented no relationships and graphs that were not integrated were coded as a different category of missing response. (Graphs that did not attempt to combine the information from the various statements were treated separately because the participant a) did not attempt to satisfy multiple constraints with their selection of data values and b) did not make use of the shared conceptual variables across the various statements.)

4.7.2 Overall Evaluation

Figures 4.9 and 4.10 show prototypical examples of both typical and atypical graphs; the features of the graphs will be discussed in more detail in the results section. A graph was defined as atypical if it was not possible to determine the dependent
variable's value. Typical graphs represented the different combinations of the variables in a representational format that enabled the determination of the DV’s values. Typical graphs could include representations of the average level of a variable, but this was not necessary.

4.7.3 Use of Graphing Conventions

Other aspects of the graph representation (e.g., format, labeling on the y-axis, and the presence of a legend) were also coded.

The format of the graph was coded as bars, lines, points, or other in terms of what was used to represent the quantities. One convention to note: in referring to the representation in general terms, the entity representing the value will be referred to as a bar. Not only is this the predominate response mode (vertical bars to represent data values), it is a conventional means for representing values on a graph. This term is used even though, as will be seen later in Table 4.5, other approaches did occur.

Labeling on the y-axis was considered in terms of whether participants explicitly indicated the DV, or whether they used implicit means (such as levels or tick marks) to indicate the presence of a DV. Additionally, it was noted whether or not the participants made use of a legend to distinguish levels of the second IV. When labels and tick marks were present on the y-axis, the participants’ choice of units and unit sizes was noted.
4.7.4 Relationships Represented

Two primary goals of this more quantitative part of the coding and analyses were to determine if a comparative statement was represented and if so, the accuracy with which the statement was represented. Thus, for instance, if the statement about the main effect of $IV_A$ indicated that $IV_{A1} = IV_{A2}$, participants’ constructed graphs were coded as to whether or not they represented this relationship and if so, whether they accurately represented this relationship. These two features of the graphs: existence of the representation and accuracy of the representation – are partially dissociable.

Given the task, it was not possible to have accuracy without also representing the existence of the relationship. However, it was possible to represent the existence of the relationship without also having accuracy (Figure 4.6, Case B). These two measures were used in order to focus on whether the manipulation specifically targeted representational performance and not the statement constraint satisfaction part of the task. It is the latter component that enables participants to determine accurate values. However, participants need to use graphing skills in order to construct the representation.
Case A
- No relationships represented

Case B
- Given statement indicating $IV_{A1} < IV_{A2}$
- Incorrectly represented data pattern in which $IV_{A1} = IV_{A2}$

Case C
- Represented all relationships accurately

<table>
<thead>
<tr>
<th>Case</th>
<th>Existence of Relationship</th>
<th>Accuracy of Relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case A</td>
<td>No</td>
<td>n/a</td>
</tr>
<tr>
<td>Case B</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Case C</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

**Figure 4.6.** Constructed graphs depicting the dissociability of the existence and accuracy of the relationship among variables. Constraints imposed by the statements determined whether a depicted representation was accurate. Here, the three selected graphs illustrate cases in which a) the relationship was not represented (and thus could not be accurate), b) the existence of the relationship was represented, but was not represented accurately, and c) the existence of the relationship was represented accurately.
Accuracy of each relationship was assessed first by direct measurement of the bar heights, and second, by coder verification of the relationship. The height of each bar was measured to the nearest half-millimeter (e.g., 2.40, 2.45, and 2.50 cm). The measured heights were used to calculate the magnitude of the relationship between the quantities indicated in each comparative statement. The actual magnitude, however, was not of specific interest. The main question was whether subjects rank-ordered the values of the variables correctly.

The comparisons relevant to the simple and main effects are the primary focus. An additional relationship was assessed when applicable. As some participants chose to represent both the overall value of the variable as well as the individual levels of a variable, these values were checked for consistency. That is, if $I_{VA1_{1}}I_{VB1}$ was represented at 5 cm and $I_{VA2_{1}}I_{VB1}$ was represented at 3 cm, was their additional representation of $I_{VA_{1}}I_{VB1}$ represented as 4 cm (the average of the two levels)? This issue, of whether participants maintained a common scale when re-representing mathematical transformations of an IV level, is addressed in Section 4.8.6.3). As this only rarely occurred, the separate representation of the main effects will not be further considered here, and the relationships are coded based on the integrated pairings of the different levels of the IVs.

The difference score was found for each of the six relevant pairings. Each difference magnitude was categorized as either small (differences < 0.4 cm) or medium/large (differences >= 0.4 cm). The medium/large differences predominated; 86% of the represented relationships (total $n = 828$) had difference magnitudes greater
than 0.4 cm. These magnitude categories were used to decide which relationships needed closer scrutiny and verification by a coder to assess the nature of the participants’ original representation of the relationship from the statement.

Most (94%) relationships of small magnitudes were checked against the graphs by both coders. Calculations indicating medium/large magnitude differences were spot-checked (26% of represented relationships). Two coders (the author and a research assistant) independently verified the nature of the represented relationship on the graph. Disagreements were resolved by a 3rd coder. After accounting for coder mistakes (10% of small magnitudes; 2% of medium/large magnitudes), agreement levels between the two coders were similar for both magnitudes of represented relationships: 87% for medium/large differences and 84% for small difference magnitudes.

Computation of the interrater reliability indicated substantial (Landis & Koch, 1977) interrater consistency for the medium/large difference sizes, Kappa= .67, SE = .05, p < .001. For the small differences, interrater reliability was only moderate with Kappa = .49, SE = .10, p < .001.

4.8 Results of Qualitative Codings of the Representational Format

Where applicable, statistical analyses are included to examine differences based on math level or condition, but for many of the qualitative codings about graph format, low frequencies prevent further statistical analyses.

---

General Notes about Results: (1) Error bars on graphs represent one standard error unless otherwise noted. (2) For repeated measures ANOVAS in which the assumption of sphericity was violated, Huynh-Feldt corrections are used if epsilon > .75, else Greenhouse-Geisser corrections are applied.
4.8.1 Evidence of Effort and Response Strategy

On average, participants took 26 (SD = 8) minutes to complete the task. Of the 5 categories considered (revision of bar heights, revision of labels, revision of graph format, reading marks, other writing), 73% of graphs showed evidence of at least one of these strategies. 31% of the graphs showed evidence of at least 2 of these strategies, and 9% of the graphs showed evidence of at least 3 of these strategies. In terms of the individual subjects, all participants showed some type of effort evidence on at least 1 of the 3 graphs they constructed. Looking at each trial individually, over two-thirds of participants produced at least one type of revision on the graph in that trial.

Across trials, revision of bar heights (39% of graphs) and revision of labels (28% of graphs) were the most common. Overall, only 11% of graphs involved both types of revision. The frequency for other effort categories was less: reading marks (27% of graphs) and writing (17% of graphs). Table 4.3 indicates that subjects more frequently produced revisions rather than reading or writing marks. In addition, the frequencies in the table indicate that reading and writing marks were roughly equally likely to be applied to 1 of the 3 trials as they were to being applied on 2 or more of the 3 trials. On the other hand, changing labels or data values was more likely to be applied on only 1 of the graphs compared to the frequency of being applied on 2 or more of the graphs.

Table 4.3. Frequencies of reading, writing, and revision (labels & data values) across participants.
Logistic regression indicated that the frequency of revising labels did decrease over time, $F(2, 85) = 7.85, p < .001$. This did not significantly interact with Condition. However, the trend was linear for the Control condition while it was at an artificially low level on the 2nd trial in the Added Labels condition because the graphs were pre-labeled. On the 1st trial, roughly half of the graphs had label revisions. Thirty percent of graphs in the Added Label 3rd trial showed label revisions, along with ten percent of the Control condition’s 3rd trial graphs. The frequency of revising data values did not change over time ($p > .05$).

### 4.8.2 Reading and Writing Marks

There was a main effect of manipulation, $F(1, 41) = 5.34, p = .02$, indicated by the logistic regression on the presence of reading marks. While the interaction between Condition and Trial was not significant, the main effect of manipulation is only present for the 2nd and 3rd trials ($p < .03$). In the Control condition, participants increased the frequency with which they produced reading marks in the later 2 trials, while participants in the Added Labels Condition did not change over time. As shown in Figure 4.7, this discrepancy lead to the result that participants in the Control condition were
more likely to produce reading marks on the 2nd and 3rd graphs than those in the Added Labels Condition.

Although the frequency of writing marks was relatively higher in the Control condition ($M = 26.7\%, SE = 6.8\%$) of graphs versus the Added Labels condition ($M = 8.7\%, SE = 6.2\%$), this difference was not significant.

![Figure 4.7](image)

Figure 4.7. Reading marks increased in the Control condition, but not in the Added Labels condition.

### 4.8.3 Overall Rates of Effort and Revision

However, the greater likelihood of applying revision techniques on only one of the three graphs (vs. applying revision techniques on two or more of the three graphs) does not come entirely from an order effect of Trial. Neither the presence nor absence of revision overall nor the number of effort categories involved were significantly affected by graph order ($ps > .05$). Still, there does appear to be a slight trend in which
the 1st graph captured the highest evidence of effort, and the 2nd graph captured the least (see means in Table 4.4).

**Table 4.4.** Revision and effort across trials. Overall amounts of revision and effort showed a trend of being highest on the 1st trial.

<table>
<thead>
<tr>
<th></th>
<th>Any Type of Revision Present?</th>
<th>Number of Effort Categories Shown</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st Graph</td>
<td>.69 (.07)</td>
<td>1.33 (.15)</td>
</tr>
<tr>
<td>2nd Graph</td>
<td>.51 (.07)</td>
<td>1.02 (.15)</td>
</tr>
<tr>
<td>3rd Graph</td>
<td>.55 (.07)</td>
<td>1.11 (.15)</td>
</tr>
</tbody>
</table>

*Note.* Mean scores are presented with standard error in parentheses. The presence of any type of revision (data values, labels, or graph format) was coded as absent (0) or present (1), and the means represent the proportion of graphs with some type of revision. The number of effort categories represents the cumulative nature of different effort categories, including reading and writing marks. This could range from 0 to 5. Logistic regression on the presence of revision and repeated measures ANOVA on the number of effort categories indicated that the effect of trial was not significant on either of these measures (ps > .05).

The total number of effort categories observed depended on a subject’s condition and their math level, but not Trial. A Math Composite (5) x Condition (2) x Trial (3) ANOVA with repeated measures on Trial, indicated that participants in the Control condition used more effort categories (M = 1.4, SE = .13) than those in the Added Labels Condition (M = .96, SE = .13), F(1, 38) = 4.48, p = .041, $\eta^2_p = .11$. In addition, participants with higher math levels used more effort categories, F(2, 38) = 7.97, p = .001, $\eta^2_p = .30$. Bonferroni adjusted follow-up t-tests indicated that the subjects at the lowest math level (M = .85, SE = .14) used significantly fewer effort
categories than those with above average math levels \((M = 1.6, SE = .14)\). The above average math group showed marginally more effort categories than the average math group \((M = 1.07, SE = .18)\). **Figure 4.8** shows the significant interaction between math and condition, \(F(2, 38) = 4.98, p = .012, \eta^2_p = .21\).

**Figure 4.8.** Effects of trial and math level on the variety of effort categories shown. In the Control condition, participants with higher math levels employed a greater number of effort categories, particularly on their 1\(^{st}\) and 2\(^{nd}\) graphs. This was not the case for low math participants in the Control condition or in the Added Labels Condition for either math group.

This qualifies the earlier main effects of Trial and Manipulation. From examining the means, it is apparent that the main effects of math background and condition as well as the interaction between the two factors are due to the high math participants in the Control condition using more effort categories on the 1st and 2nd graphs. Math Background and Condition did not influence whether or not any revision was shown.
4.8.4 Representational Format of Graph

Figure 4.9 presents examples of traditionally represented graphs, and Figure 4.10 presents examples of atypical graphs. An important note applies to the descriptions of the atypical figures. Although an attempt is made to describe how the representation might relate to the relationships indicated in the statements, these are explanations generated by the author and colleagues studying the produced patterns. Although the strategy generates a pattern that works with what the participant represented, they are hypothetical strategies, and do not necessarily represent the participants’ process.

Roughly half of the graphs were atypical in some way, even if they still contained traditional features. Less severe ‘violations’ include representing the independent variables as being on a continuous scale, connecting lines between the values, and adding additional columns. These are considered less severe because each of these ‘violations’ are formats that can be used appropriately. But, as demonstrated by these participants, the formats can also be used inappropriately. More severe violations include placing one IV on the x-axis and the other IV on the y-axis, not including the dependent variable, not basing the graph on the data relationships, and not integrating the statements – especially if the represented values were inconsistent with one another. Note that these categories are violations of what would be appropriate graphing practices. Participants often made more than one type of error on each graph. These patterns are covered in this Section (4.8.5) and the following Section (4.8.6) on the use of graphing conventions.
Figure 4.9. Examples of typical graphs. Across these various representations, participants fully crossed the two IVs, thus representing all 4 quantities. However, it was also possible to have nontraditional features combined with otherwise typical representations. For instance, the constructed graph in 511-10 connected the magnitudes in a curvilinear fashion. Other aspects to note include the unique use of the legend key in 527-13, and the presence of data revision in 507-13. These examples all labeled the DV, although that was not necessarily the case across subjects.
Figure 4.10. Examples of atypical graphs.

**535-10** starts as a traditional graph, with its representation of the crossed IVs of danger level and bystander presence in the first two pairs of bars. Then, it starts to become inconsistent as it represents the two IVs without their levels, and it shows a greater level of confusion as it mixes in the DV as category labels for the last set of bars.

**512-5** is interesting in that each IV is represented with its two levels, and two levels of the DV are represented as well. The difficulty comes when trying to determine how the variables relate to one another. One possible interpretation of the graph is as a stream-of-consciousness of the participant reading the statements (e.g., the first statement indicated that ‘low danger’ ‘helped less’. However, this was not necessarily the participant’s strategy.

**505-5** presents an example of a graph with multiple violations. The two IVs are made into continuous variables (along with additional levels that were given data), and the DV
is possibly represented in a 3rd dimension as marked by the words. The graph is also atypical in that it does not integrate the statements. Although it becomes less clear when there are multiple lines for a given value of sleep, part of the participant’s strategy might have been to take each statement and indicate which level of sugar represented more and which was less.

532-5 The strategy of this participants’ response cannot be determined. Of note, one IV was on each axis. However, it is also critical to note that the concept described in words referenced the idea of interacting variables. Their description, however, is not represented in the graph. They did correctly describe the main effect of IV_A: people who were alone helped more than people who were accompanied.

Bar graphs predominated as response types. Classifying subjects by the format of the graph they used on at least 2 of the 3 trials, 84% of the subjects created bar graphs, as seen in Table 4.5. The remaining subjects were evenly divided among line graphs, point graphs, and atypical representational formats. Looking just at the graphs themselves, 77% of the formats were bar graphs.

Table 4.5. Bar graphs predominated as response types.

<table>
<thead>
<tr>
<th></th>
<th>Bar</th>
<th>Line</th>
<th>Point</th>
<th>Point &amp; Line</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>By Subject (at least 2 of their 3 graphs)</td>
<td>84%</td>
<td>2%</td>
<td>4%</td>
<td>4%</td>
<td>4%</td>
</tr>
<tr>
<td>By Graph</td>
<td>77%</td>
<td>4%</td>
<td>4%</td>
<td>7%</td>
<td>7%</td>
</tr>
</tbody>
</table>

Looking at the three representations by subject, 64% of subjects used the same graph format on all three trials; the remaining subjects \( n = 16 \) used the same graph format on 2 of the 3 trials. Of these 16 subjects, 10 of the 11 in the Added Labels Condition produced the same format on the 2nd and 3rd graph. Of the 5 subjects that were in the Control condition, they were divided among pairing the 1st and 2nd graph
and the 2\textsuperscript{nd} and 3\textsuperscript{rd} graph. Thus, participants in the Added Labels Condition switched representational formats and maintained it on the 3\textsuperscript{rd} graph.

4.8.4.1 Traditional graphs

Participants’ graphs were also scored as to whether or not they represented a traditional or conventional approach of graph construction. Criteria for this were relatively lenient. There had to be 4 identifiable data values, but the data values could be grouped together or not, and there could also be additional data values present. The DV did not need to be labeled, and any means of identifying the 4 values arising from the crossed IVs was acceptable.

The frequency of traditional representations was assessed with a Condition (2) x Trial (3) x Composite Math (3) logistic regression with repeated measures on Trial. This yielded a marginally significant main effect of math level on the frequency of traditional graphs overall, $F(2, 10) = 2.93$, $p = .06$. Comparisons of the log odds estimates indicated that indicated that the lowest math level ($M = 50\%$, $SE = 9.6\%$) produced significantly fewer traditional graphs than the highest math level ($M = 82\%$, $SE = 10\%$). In addition, the analyses yielded a main effect of trial, $F(2, 85) = 5.55$, $p = .005$. On the first trial, 50.1\% ($SE = 9.8\%$) of the graphs were traditional. The second trial elicited 74.7\% ($SE = 7.1\%$) traditional graphs, and the 3\textsuperscript{rd} trial elicited 67.2\% ($SE = 7.2\%$) traditional graphs. Pairwise contrasts showed that the third trial had significantly more traditional graphs than the first trial, and significantly less than on the second trial ($p < .05$, Bonferroni-adjusted).
Of more interest here, however, is the significant Trial x Condition interaction, $F(2, 85) = 3.22, p = .04$. Simple effect analyses showed that the effect of Trial was only present in the Added Labels condition, $F(2, 85) = 8.54, p < .001$. As shown in Figure 4.11, the frequency of traditional graphs increased on trial 2 in the Added Labels Condition, and this increase was maintained at trial 3. A priori pairwise contrasts indicated that Trial 1 was significantly less than both Trial 2 and Trial 3 ($p < .002$), but Trials 2 and 3 were not significantly different from one another, $p > .3$. There was no change in the use of traditional graphs in the control group, $F(2,85) = 0.17, p = .8$. In addition, on the first trial, participants in the Added Labels condition produced marginally fewer traditional graphs than participants in the Control condition, $F(1, 85) = 3.71, p = .06$.

**Figure 4.11.** Traditional graphs were more frequently represented during and after the manipulation in the Added Labels condition. No such effect was found for the Control condition.
Focusing on subjects in the high math group and in the low math group (leaving out the average group of \( n = 7 \)), it is possible to ask whether these gains from the Added Labels condition occurred at each math level. As predicted a priori by the hypotheses about the manipulation, subjects in the Added Labels condition had significantly more graphs on the 2nd trial than on the 1st trial, \( p = .03 \). Trials 2 and 3 did not differ, \( p > .4 \), but Trail 1 was marginally lower than Trial 3, \( p = .08 \). The subjects in the higher math group also demonstrated a gain from the 1\(^{st} \) (\( M = 58\% \), \( SE = 13\% \)) to 2\(^{nd} \) (\( M = 92\% \), \( SE = 12\% \)) trial (\( p = .057 \)), and they maintained the same level of performance on the 3\(^{rd} \) trial.

### 4.8.4.2 Graphs with one IV per axis

Figure 4.12 shows examples of participants’ graph constructions when only 1 IV was used on each axis (and the DV was not represented on the graph). The low frequency of this type of response precludes statistical analysis. Across trials, participants decreased the frequency with which they produced these atypical representations of one IV per axis. On the 1\(^{st} \) trial, eight subjects produced graphs with one IV on each axis. Two of these were in the Control condition – one continued to make the same error on the subsequent graphs, the other did not. Of the six that were in the Added Labels condition, none of them made the error on their subsequent graphs. Thus, it appears that the Added Labels condition could have been beneficial in moving subjects away from this representational error.
Figure 4.12. Examples of constructed graphs in which 1 IV was placed on each axis. These two graphs are also atypical in others ways. 545-18 represents continuous dimensions. In addition to adding multiple levels, 505-12 does not make use of the data. This latter graph possibly resembles an attempt at a prototypical multi-variable bar graph.

4.8.4.3 Integrating the data values

Figure 4.13 depicts an example of a non-integrated graph. Other versions of constructed graphs in this category used the labels from the text, but did not attempt to combine the statements. Four subjects produced non-integrated graphs on the 1\textsuperscript{st} trial. All of these subjects were in the Control condition. Three of them continued to use the same approach on their subsequent trials. The remaining subject switched to an integrated representation for the 2\textsuperscript{nd} and 3\textsuperscript{rd} trials. Only one subject produced a non-integrated graph after initially producing an integrated graph.
Figure 4.13. Example of a non-integrated graph. Of note here, in example 538-10, is the fact that the legend did not change when the participant started the second set of statements – even though the variable being compared in the statements had changed. In addition, lack of integration is seen by the representation of the alone bar of statement 1 as greater than either the alone bar of statement 2 or the alone bar of statement 3. (Statement 1 describes the average of the values in statements 2 and 3.)

4.8.5 Tick Marks and Unit Labeling on the Vertical Axis

As the instructions specifically told participants that it was not necessary to use units on the DV, it is not surprising that few of them used tick marks and/or units on the DV axis. Table 4.6 indicates the prevalence of these categories. Of note is that fact that when participants did give the DV a scale, they represented it using equal unit sizes. In fact, only one participant, on a single graph, had unequal units, and the representation was also atypical in other ways. The 4 graphs that used equally spaced tick marks with non-numerical labels used relatively scaled measurements that divided up the axis into levels (e.g., low, medium, and high).
Table 4.6. Frequency of tick mark usage and unit spacing on the y-axis. In this task, participants rarely explicitly indicated the units or unit size on the y-axis. However, when participants did so, they used equal units. (Graphs counted here may have other atypical features (e.g., 1 IV per axis in the labeled endpoints category).

<table>
<thead>
<tr>
<th>Tick Mark Placement</th>
<th>Equally Spaced Tick Marks</th>
<th>No Tick Marks</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>With Tick Mark Labels</td>
<td>With Tick Mark Labels</td>
</tr>
<tr>
<td></td>
<td>Numbered Tick Mark Labels</td>
<td>Labeled endpoints (or endpoints + center)</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>With No Tick Mark Labels</td>
</tr>
<tr>
<td>Number of Instances</td>
<td>14 graphs, across 6 subjects</td>
<td>5 graphs, across 5 subjects (7 graphs, across 6 subjects)</td>
</tr>
<tr>
<td></td>
<td>4 graphs, across 2 subjects</td>
<td>Remainder of Ss</td>
</tr>
<tr>
<td></td>
<td>6 graphs, across 4 subjects</td>
<td>2 graphs, across 2 subjects</td>
</tr>
</tbody>
</table>

4.8.6 Representing IV Levels and Averages

While the use of equal sized units is important for comparing heights of unrelated bars, it becomes even more critical when you are representing data values that share components (e.g., representing the individual values for the levels as well as the average value). Thirteen graphs (across 9 subjects) represented both individual values as well as average values. Of these, 2 subjects (3 graphs total) divided the DV axis into units, 1 subject labeled endpoints without using tick marks, and 1 subject used numbered tick marks on their DV axis. The others did not use tick marks or units. Three other subjects (4 graphs total) represented only the average values of the IVs and not the values for the individual levels. These counts are among participants who
represented at least some of the values in a category, but they did not have to represent all of them to be considered here.

Thus, representing average values was relatively rare. Examining each graph as a whole, the subjects did not tend to represent the average as an un-weighted (or even as a weighted) average of the individual levels. That is, while occasionally the values would be consistent, other values were not even in the appropriate direction (e.g., the average was half the value of the two individual levels). This inconsistency in the representation of both individual values and average values for IVs marks a lack of understanding of the relationship between these the average value and individual values for the IVs. However, as representing the averages was not a necessary component of the task and as it rarely occurred, this is an issue that needs further investigation in future research.

4.9 Results: Representing Relationships

In this construction task, it is possible to represent the existence of a relationship on the graph without representing the relationship accurately. Thus, three different measures will be considered: existence (the proportion of relationships represented), raw accuracy (the proportion of the 6 relationships correctly represented), and adjusted accuracy (the proportion of represented relationships that were correctly represented).

4.9.1 Existence of Relationships on Graph

If the constructed graph represented at least one relationship, it represented all six of the relationships among the variables 90% of the time. Graphs representing only some of the relationships include responses where the variables were not fully crossed
as well as those few that crossed the variables but did not assign magnitudes to all the combinations. Over half (53%) of all graphs represented all the relationships all of the time; this was 69% of the subjects. Twenty-five graphs (across 16 different participants) represented no relationships. As seen in Figure 4.14, the majority of these graphs were on the 1st trial.

![Figure 4.14](image)

**Figure 4.14.** Graphs representing all, some, or no relationships by trial. Each bar represents all the graphs constructed on a given trial. The lowest section in green, indicates the frequency of graphs representing all of the 6 relationships (the main and simple effects for each IV). The middle section in red represents graphs representing some, but not all, of the relationships. The top section, in blue, represents graphs that did not represent the existence of any of the relationships. The proportion of graphs representing none of the relationships was highest on the 1st trial.

The focus will be on the effects present within each condition. A significant Trial x Condition x Math interaction, $F(3.2, 59.9) = 3.78, p = .01, \eta_p^2 = .17$, in the overall analysis (along with lesser effects) prompted two follow-up ANOVAs, one per condition, to assess the percentage of relationships represented in each graph.
The effect of Trial (3) and Math (3)\textsuperscript{29} were assessed within each condition, with repeated measures on Trial. In the Control condition, there was no main effect of Trial, $p = .5$, although the Trial x Math interaction was marginally significant, $F(2.1, 17.2) = 3.49, p = .053, \eta_p^2 = .29$. This arises from the lower math group increasing their proportion of relationships represented from the $1^{\text{st}} (M = .38, SE = .15)$ to the $2^{\text{nd}}$ and $3^{\text{rd}}$ trials ($M \sim .67, SE = .15$). However, this is not significant once significance levels are corrected for multiple comparisons.

In the Added Labels condition, the ANOVA yielded a main effect of Trial, $F(1.8, 38.1) = 10.41, p < .001, \eta_p^2 = .33$ (see Table 4.7). Pairwise comparisons (with $p$ values Bonferroni adjusted) indicated that participants represented more relationships on the $2^{\text{nd}}$ graph than they did on the $1^{\text{st}}$ graph ($p = .002$). There were fewer relationships represented on the $3^{\text{rd}}$ graph than on the $2^{\text{nd}}$ graph ($p = .023$). However, while the mean for the $3^{rd}$ graph was still higher than on the $1^{st}$ graph, it was not significantly different ($p = .09$).

Table 4.7. Proportion of relationships represented, regardless of accuracy. Participants significantly increased from the $1^{\text{st}}$ to $2^{\text{nd}}$ trial in the Added Labels condition. However, they did drop significantly from the $2^{\text{nd}}$ to $3^{\text{rd}}$ trial, which was not significantly different

\textsuperscript{29} Using a different math division, the composite math measure was divided into two groups (low: $n = 16$; high: $n = 28$). A Condition (2) x Math (2) x Trial (3) ANOVA with repeated measures on Trial yielded a main effect of Trial, such that the $1^{\text{st}}$ trial represented fewer relationships than the subsequent trials, $F(1.5, 61.2) = 6.76, p = .005, \eta_p^2 = .15$, pairwise comparisons with LSD adjustments $p < .04)$. In addition, the interaction between Trial x Condition x Math was significant, $F(1.5, 61.2) = 3.95, p = .035, \eta_p^2 = .09$. Lower math participants increased the number of relationships represented in the Control condition as well as in the Added Labels condition from the $1^{\text{st}}$ to the $2^{\text{nd}}$ graph. They did not have further increases going into the $3^{\text{rd}}$ graph. The higher math participants in the Added Labels condition increased from the $1^{\text{st}}$ to the $2^{\text{nd}}$ graph and maintained that increase for the $3^{rd}$ graph.
from the first trial. Trial did not affect the representation of relationships in the Control condition.

<table>
<thead>
<tr>
<th></th>
<th>1st Graph</th>
<th>2nd Graph</th>
<th>3rd Graph</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control Condition</td>
<td>.70 (.10)</td>
<td>.75 (.10)</td>
<td>.76 (.10)</td>
</tr>
<tr>
<td>Added Labels Condition</td>
<td>.55 (.10)</td>
<td>.96 (.05)</td>
<td>.78 (.07)</td>
</tr>
</tbody>
</table>

Looking at the differences within each math group for the Added Labels condition (see Figure 4.15), clear patterns are apparent despite the lack of significance. The increase and subsequent decrease in the lower math group did not reach significance. The average math group significantly increased from Time 1 (Bonferroni adjusted \( p = .01 \)) to Time 2, but the subsequent decrease resulted in Time 3 performance that was not different than that of Time 1, once adjusted for multiple comparisons. In the higher math group, the participants increased to ceiling (\( p = .17 \)), and thus could not show further improvement. The higher math group remained at ceiling on the 3\(^{rd} \) trial. Thus, the degree to which the gain that participants in the Added Labels condition demonstrated in the 2\(^{nd} \) graph transferred to the 3\(^{rd} \) graph depended on math level. Participants in the Control condition did not have a similar increase in representing the existence of relationships.
**Figure 4.15.** Representing the existence of relationships increased in the Added Labels condition, but not the Control condition. Overall, participants in the Added Labels condition represented significantly more relationships on the 2nd than on the 1st trial. This gain was maintained for the higher math group, and partially maintained for the average math group. There was no effect of Trial in the Control condition, although the lowest math group did significantly increase from Time 1 to Time 2.

### 4.9.2 Accuracy of Relationships on Graph

As existence was already considered above, the focus here will be on the adjusted accuracies, that is, the accurate depiction of the comparative relationship specified in the statement, given that the relationship was represented. The sample size ($n = 29$) was lower for these analyses as participants had to represent at least one of the possible relationships on each trial to be included.

To assess the effect of trial, manipulation, and math on the adjusted accuracy scores, a Trial (3) x Condition (2) x Math (3) ANOVA was run. There was a main effect
of trial, $F(2, 46) = 4.36, p = .018, \eta^2_p = .16$. Bonferroni adjusted $p$ values in the post hoc comparison of the levels of trial indicated that accuracy on the 3rd graph was marginally higher than on the 2nd graph (adjusted Bonferroni $p = .07$). (See Table 4.8 and Figure 4.16.) As accuracy rate was not affected by labeling condition, Figure 4.16 presents data that collapses across condition. Accuracy on the pre-test graph and the post-test graph did not differ from one another – again, this is within the set of graphs that did represent the relationships.

There was a main effect of math, $F(2, 23) = 5.37, p = .012, \eta^2_p = .32$. Participants in the lowest math group ($M = .78, SE = .03$) represented the relationships significantly less accurately than participants in either the average ($M = .92, SE = .04$) or higher math groups ($M = .89, SE = .03$), Bonferroni-adjusted $ps < .03$.

### Table 4.8. Proportion of relationships represented accurately (out of all 6 relationships).

<table>
<thead>
<tr>
<th></th>
<th>1st Graph</th>
<th>2nd Graph</th>
<th>3rd Graph</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control Condition</td>
<td>.88 (.04)</td>
<td>.81 (.06)</td>
<td>.91 (.04)</td>
</tr>
<tr>
<td>Added Labels Condition</td>
<td>.87 (.04)</td>
<td>.79 (.05)</td>
<td>.94 (.03)</td>
</tr>
</tbody>
</table>

Added Labels condition, $F(1.9, 35.6) = 3.98, p = .030, \eta^2_p = .17$. However, pairwise comparisons indicated that no means were significantly different from one another (adjusted Bonferroni $ps > .1$). No other effects were significant for either condition.
Overall, the adjusted accuracy rate for the Control condition and Added Labels condition were the same. This does not negate the previously presented result that more relationships were represented in the Added Labels condition as it focuses on accuracy within the represented relationships. Another way of stating this is that even though more relationships were represented in the Added Labels condition, there were similar accuracy levels in the two conditions for the represented relationships ($p = .19$).\footnote{This does imply, however, that the number of correct representations was higher in the Added Labels condition because the Added Labels condition had more graphs that had measurable representations. This issue is independent, however, from whether participants could accurately represent the relationships given that they produced a typical graph format. This issue of accuracy of represented relationships was not affected by the Added Labels condition.}
Figure 4.16. Accuracy of represented relationships across trial and math level. Graphs 1 and 3 did not differ in their accuracy levels. Accuracy (among graphs which represented the relationships) was not affected by the manipulation in any of the three math levels. Accuracy was lower on the 2nd trial. Lower math had significantly lower accuracy levels than average or higher math.
4.9.3 Accuracy of Relationships based on Statement

In addition to the above analyses, which treated each graph as the unit of analysis, it is also possible to ask about the differences of accuracy across the different statements. If you consider all of the graphs (including those that were atypical and thus could not accurately represent the relationships), the raw accuracy of the graphed relationships for the main and simple effect statements ranged from 49 – 84%. When focusing only on relationships that were represented, the adjusted accuracies ranged from 59 – 100% across statements (\( M = 86\%, SD = 12\% \)) per graph.\(^{32}\)

Statement accuracy on a given trial is either correct or incorrect. Participants were most accurate on simple effect statements. In particular, the second simple effect statement for IV\(_A\) and the first simple effect statement for IV\(_B\) were represented more accurately than the main effect statement for IV\(_B\) and the other simple effect statement for IV\(_B\). Main effects were graphed accurately around 60% of the time.

This patterning of differences (as seen in Table 4.9) may reflect a tendency to deal with the statements in the presented order. Perhaps after satisfying one of the simple effects for an IV, it was more difficult to also satisfy the other simple effect. This might have been more pronounced for IV\(_B\) as those statements were presented after the statements about IV\(_A\). In addition, the main effect statement for IV\(_A\) was significantly lower than the first simple effect statement for IV\(_A\).

\(^{32}\) Few graphs represented some, but not all, of the relationships, so this did not have a large effect on the range of accuracy levels. If considering only the graphs that represented at least some of the relationships and counting missing relationships as incorrect, accuracies ranged from 58 – 95%.\[\]
Table 4.9. Math level and statement type both impacted the accuracy of performance. These analyses are based only on graphs that represented relationships. The level of performance seen in the lower math group was less accurate (as seen by both a larger number of ‘harder’ statements and a lower average accuracy of harder statements).

<table>
<thead>
<tr>
<th>IV</th>
<th>Statement</th>
<th>Below Average Math</th>
<th>Average Math</th>
<th>Above Average Math</th>
</tr>
</thead>
<tbody>
<tr>
<td>IVₜₐ</td>
<td>Main effect</td>
<td>Harder</td>
<td>Harder</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Simple effect at IVₜₐ₁</td>
<td>Harder</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Simple effect at IVₜₐ₂</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IVₜₐ</td>
<td>Main effect</td>
<td>Harder</td>
<td>Harder</td>
<td>Harder</td>
</tr>
<tr>
<td></td>
<td>Simple effect at IVₜₐ₁</td>
<td></td>
<td></td>
<td>Harder</td>
</tr>
<tr>
<td></td>
<td>Simple effect at IVₜₐ₂</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*M and SD of ‘Harder’*  
- Below Average Math: .71 (.06)  
- Average Math: .77 (.03)  
- Above Average Math: .79 (.03)

*M and SD of ‘Easier’*  
- Below Average Math: .97 (.04)  
- Average Math: .99 (-.02)  
- Above Average Math: .95 (.03)

4.10 Summary and Discussion

Despite the high demands posed by this graph construction task, participants who demonstrated an understanding of graphical conventions were relatively successful at satisfying the constraints imposed by the statements and accurately representing the relationships. However, the qualification of “demonstrated an understanding of graphical conventions” is important to note. This graph construction task presented considerable difficulty for some of the participants. Regardless, participants showed that they attempted the task. The majority revised labels at least once, and they also revised data values at least once.

Providing participants with a pre-labeled graph appears to have greatly aided their ability to generate typical representations, as measured by both the qualitative
and quantitative coding schemes. The only difference introduced in this manipulation was the addition of labels to conventional locations on the graph. Of those who started out with atypical representations, those in the Added Labels condition switched to more typical representational formats on the 2nd trial. Aspects of this (traditional format, bar graphs, legends, DV) tended to carry over to their 3rd trial as well. No such gain was seen in the Control condition.

Assessing the proportion of the six relationships represented in the constructed graph is another approach to measuring the typicality of the graph representation. This measure of typicality, the representation of main and simple effect relationships, allows extra columns (even inconsistent extra columns or those inappropriately mixing in the DV) without penalty and does not assess the typicality of other aspects of graphing conventions. Instead, it focuses on whether or not the data could be obtained from the graph – a question integrally tied to the communicative function of a graph.

In terms of the relationships represented, in the Added Labels condition, they represented more of the relationships on the 2nd than on the 1st trial. Those in the Control condition did not change overall, although there was some indication that there was a beneficial practice effect for those in the lower math groups. (Random assignment led to the relatively lower level of representation in the Added Labels condition on the 1st trial.) The transfer of the gain in the Added Labels condition depended on the level of math background, with more math background being positively associated with transfer.
Separating out representational existence from representational accuracy is another way of separating the statement constraint satisfaction component of the task from the graphical representation component of the task. This focuses on the idea that the process (formatting decisions, representation of relationships) is different from the product (the accuracy of the represented relationships). It was found that adding graph-relevant labels to the graph representation improved representational format, but not accuracy. This draws attention to the component processes that contribute to graph construction. The representational components of the task were differentiated from the constraint satisfaction and the accuracy of dealing with the relational comparison statements. However, successful completion of the constraint satisfaction aspect required a minimal amount of representational skills.

Successful resolution of the constraints can only be demonstrated if the generated values could be placed on the graph. An additional component here is knowing that the task is requesting 4 values and how the variable levels should be assigned to those values. One broad category of errors seen in the presentation of the qualitative results relates to the role of the variables.

One of these types of error patterns seen was the assignment of one independent variable to each axis (section 4.8.4.2). Other errors included mixing the IVs and the DVs along the x-axis. A third error type related to working with the variables was incomplete crossings of the two IVs. That is, participants may have initially represented the category level of one IV before then representing one of the value levels as well. This is related to a 4th recurring error in variables in which the average of
the two levels for an IV was represented in addition to the two levels, but the represented average did not equal the mathematical average. This type of error could occur if the participants lacked the recognition that the two levels are what determine the overall category of the variable. These differences were evaluated by the measurement and coding scheme described in 4.7.4. Thus, these errors related to spatially locating the IVs and DV on the graph representation could arise at the representational level, but they could also arise earlier in participants’ conceptualization of variables.

Overall, though, this is evidence supporting the idea that one of the difficulties participants encounter when constructing graphs is spatially locating the independent and dependent variables on the axes. This, combined with the large increase in traditional formats and existence of relationships in the Added Labels condition, supports the proposed theoretical role of the manipulation. That is, it supports the idea that the manipulation is helping participants structure how the variables are related to one another and/or the graph. This could be either a reminder or as a guide that leads participants to implement greater structure and representational conventions.

Different levels of performance on the 1st graph, combined with varying math levels, seem to obscure some of the effects. For instance, whether or not the benefit received on the 2nd graph in the Added Labels condition carried over to the 3rd graph varied based on how math background was operationalized. In addition, its transfer appeared more prevalent among high math participants than low math participants. Math appeared to impact the accuracy of subjects in the Control condition, but not the
accuracy of subjects in the Added Labels condition. Math also differentially affected the number of effort categories demonstrated; high math subjects in the Control condition demonstrated more effort categories on the 1\textsuperscript{st} and 2\textsuperscript{nd} trials than other subgroups of participants. This appears to have arisen from a larger amount of reading and writing marks.

Overall, accuracy levels of represented relationships were not affected by the positional order of the graph or by condition. This general statement is qualified by noting that the 2\textsuperscript{nd} trial tended to have lower accuracies. However, the similar accuracies on the 1\textsuperscript{st} and 3\textsuperscript{rd} trials (on which the data patterns were counterbalanced) suggest that this is not an issue of the positional order, but rather an issue of the graph pattern. In general, however, there was not a practice effect on participants’ ability to complete the constraint satisfaction component of the task, nor did there appear to be a large practice effect on the ability to generate a typical graph. The lack of a change in accuracy among typical graphs suggests that the constraint satisfaction aspect was a lesser challenge than generating a graphical representation. This is surprising given the predicted difficulty of satisfying the statement constraints. In particular, the listed order of the statements (with main effects first) adds more difficulty than if one were to start with graphing the simple effects. The actual processes that participants undertake to satisfy these statement constraints can be studied in future research by having participants think aloud as they work. This same method would allow an investigation of how subjects deal with the redundancy, that is, that some statements expressed information that was also included in another statement constraint.
Accuracy did depend on statement type, however. In Section 4.9.3 it was mentioned that only one of the simple effects for each IV elicited significantly more accurate representations of the relationship. In addition, the other statements about IVB were lower than the remaining statements. This patterning of differences may reflect a tendency to deal with the statements in the presented order. Perhaps after satisfying one of the simple effects for an IV, it was more difficult to satisfy the other simple effect. This might have been more pronounced for IVA as those statements were presented after the statements about IVB. In general though, the statements about main effects were represented less accurately than the statements about simple effects.

Participants were not explicitly told that the data referenced by the comparative statements came from the same sample size for each group. Having equal sample sizes is a necessary criterion for averaging the main effect from the individual levels; otherwise, a weighted average process is needed. The limited number of participants who explicitly represented the average value of each IV (i.e., collapsing across the IV’s two levels) reduces the impact of this, but it would be information to include in future studies of this type.

Future work addressing graph construction should note the impact of the data pattern on the accuracy of the represented relationships. While this was not fully analyzed here due to small sample sizes, examining the mean accuracy levels for the different data patterns is suggestive of differential impacts of the different graph patterns. Table 4.10 summarizes the complexity levels that each pattern represented; the patterns themselves were shown in Figure 4.2. It appears that participants found it
more difficult to satisfy the constraints imposed by the graph patterns on the 2nd trial (patterns 12 and 13) relative to the patterns used in the 1st and 3rd trials. This is reflected by the lower accuracy levels on the 2nd graph, as shown in Figure 4.16. In addition, for the participants who received Set B (patterns 10, 13, and 20), it was easier to meet the constraints for pattern 10, regardless of whether it came 1st or 3rd. For participants receiving Set A, pattern 5 had a slight tendency to be easier than Pattern 18. These differences are consistent with the theoretical ideas behind the simple effect consistency model in that data patterns at higher complexity levels will be more difficult to use.

<table>
<thead>
<tr>
<th></th>
<th>IV&lt;sub&gt;A&lt;/sub&gt; Complexity Level</th>
<th>IV&lt;sub&gt;B&lt;/sub&gt; Complexity Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set A Pattern 5</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Set A Pattern 18</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>Set B Pattern 10</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Set B Pattern 20</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Set A Pattern 12</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>Set B Pattern 13</td>
<td>3</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 4.10. Performance varied based on the graph pattern. Of those patterns presented on the 1st and 3rd trials, the less complex pattern tended to be easier, regardless of whether it was presented 1st or 3rd. Patterns on the 2nd trial were (patterns 12 & 13) were more difficult overall. One hypothesis is that the lack of symmetry interferes with interpretation – perhaps because of incorrect interpretation skills.

The counterbalancing of the graph patterns in the 1st and 3rd trials (along with using both Set A and Set B of patterns) reduces the possibility that practice or fatigue
drive these differences. Instead, satisfying the combination of constraints indicated by the statements for a particular pattern appears to play a role in representational accuracy. This requires a continuation of the study beyond the data available here, as the graph patterns are collapsed in the analyses to examine the effects of the Added Labels manipulation.
CHAPTER 5

EFFECT OF STATISTICS EDUCATION ON GRAPH INTERPRETATION

5.1 Overview

In the graph construction task (Chapter 4), the accuracy with which participants represented the relationships among the variables was an indirect measure of their interpretation abilities. The construction task also assessed representational components of graphical literacy as well as requiring constraint satisfaction skills. Overall though, participants had more difficulty accurately representing the main effects than the simple effects. The study presented in this chapter looks at interpretative competence more directly by asking subjects to evaluate various types of interpretation statements in reference to presented graphs.

One of the functions of text material that typically accompanies graphs (either as part of the graph or as a description of what is shown, see section 1.7) is discussing how the variables relate to one another. Titles and captions do have a clear purpose: clarifying the information communicated by the graph. Other text can integrate information about the different IVs and the DV and focus on how the variables relate to one another. It is these types of interpretation statements - statements that examine how variables relate to one another and the DV – that are part of the present investigation.

Undergraduate students evaluated whether text statements accurately reflected data patterns depicted in a target graph at the beginning and end of a statistics course.
For example, Figure 5.1 shows a sample statement with the four accompanying graphs. Three main questions were addressed in the study:

- How well can undergraduates evaluate the accuracy of a statement about a graph?
- What is the effect of a statistics course in psychology on graph interpretation?
- How is statement evaluation influenced by variations in graph pattern?

Given the literature, participants were expected to have higher accuracy levels when evaluating simple effects, lower accuracy levels when evaluating interaction statements, and intermediate accuracy levels for statements about the values of the main effects of the IVs. The task was administered at the start and end of a basic statistics course in psychology. One focus of the statistics course was on understanding group means and basic inferential statistics (including ANOVAs). It was expected that this training would transfer to graphical understanding. Specifically, Time 2 scores were expected to be higher than Time 1 scores, especially for interpretation statements about interactions.

The effect of word meaning on the number of words recalled depends on the type of distraction.
Figure 5.1. Sample problem from Time 2 material. Each of the 6 statements in the task was accompanied by 4 graph patterns. Participants indicated whether or not a statement (seen in box above the 4 graphs) was true for each presented graph.

5.2 Participants

A total of 45 undergraduates (Age: $M = 21.7$, $SD = 5.33$, $Mdn = 20$; Female = 71%) from two summer psychology statistics courses participated at both time points: the first day (Time 1) of the course and during the last week (Time 2) of the course. The gender ratio in the sample reflects the make-up of the classes; most students in the classes chose to participate. Almost half of the participants were at least in their 4th undergraduate year, reflecting a strong tendency to delay taking this required course. Participants attended a large northeastern university, known for its socioeconomically and ethnically diverse student body. An additional 9 subjects participated only at Time 1, and another 9 subjects participated only at Time 2. Another student was excluded from data analysis for failing to complete the task at Time 2.

Participants spent 10-20 minutes completing the questionnaire during a break in their class. Students were offered extra credit and/or enrollment in a lottery for a movie ticket (odds approximately 1:20). All subjects signed informed consent forms.

Participants were split into two levels of math experience based on their highest level of math taken. The higher group consisted of those who were currently taking, or had previously taken, Calculus I; of these 17 participants, only 3 had Calculus II or more. In the lower math group ($n = 28$), 68% of this group indicated that they were currently taking, or had previously taken, pre-calculus as their highest level of math taken.
Level of math experience, previous statistics experience, math SAT, and major were also positively correlated with one another [math experience & SAT: \( r(32) = .60, p < .001 \); SAT & previous statistics: \( r(32) = .38, p = .026 \); math / science major & previous statistics: \( r(34) = .52, p = .001 \); math experience & previous statistics: \( r(35) = .40, p = .015 \)]. These correlations were stronger and more prevalent among these variables when the whole sample was included. Binary levels of math experience, major, and math SAT were combined to form 3-level mean-centered composite variable (below, at, or above average math background).

Thirty percent of the students had previously taken a statistics course; the majority of these students were in the higher math group. The statistics courses in which this study was conducted covered ANOVAs (including interactions) just before Time 2, and the courses’ final exams covering this material were within a couple of days.

### 5.3 Stimuli

#### 5.3.1 Statements

Participants read a total of 6 statements in this task. The example problem presented earlier in Figure 5.1 included one of these statements. As shown in Table 5.1, the statements can be divided into the categories of simple effects, main effects, and interactions. Both the simple effects and the interaction statements had directional and non-directional versions. The directional version specified how the variables related to one another; the non-directional version simply specified that the variables were related. Table 5.1 also gives the syntactic versions of the six statements (Section 1.3.2
and gives an overview of this notation (see also Box 6.2). Time 1 and Time 2 used similar statements – the only difference was the variable names because of the different content of the graphs; this is evident in the shared syntactic notation for both timepoints. The 6 statements were chosen such that some responses would be true, while others would be false. All statement types were presented as T/F statements. Across all statements, true was the correct response for 67% of the responses.
Table 5.1. Six statements were presented to participants at each time in this task. The wording of the statements is presented below along with a description of the type of statement and the syntactic version. Time 1 and Time 2 Statements differ only in terms of the content, as seen by the substitution of different variable names.

<table>
<thead>
<tr>
<th>Type of Statement</th>
<th>Nature of Relationship in Statement</th>
<th>Statements Seen by Participants at Time 1</th>
<th>Statements Seen by Participants at Time 2</th>
<th>Syntactic Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple Effect</td>
<td>Directional</td>
<td>With big groups, subjects donated more with strangers than with friends.</td>
<td>With different words, subjects recalled more words when hearing a podcast than when hearing a talk show.</td>
<td>In $IV_{A2}$, $IV_{B1} &gt; IV_{B2}$</td>
</tr>
<tr>
<td>Simple Effect</td>
<td>Non-Directional</td>
<td>In little groups, the amount donated was not affected by the familiarity of the group members (i.e., strangers or friends).</td>
<td>With related words, the number of words recalled was not affected by the type of distraction (i.e., podcast or talk show).</td>
<td>No effect of $IV_{B}$ at $IVA_{A1}$</td>
</tr>
<tr>
<td>Main Effect</td>
<td>Non-Directional</td>
<td>Little groups and big groups had the same average donation amount.</td>
<td>Related words and different words had the same average number of words recalled.</td>
<td>$IVA_{1} = IVA_{2}$</td>
</tr>
<tr>
<td>Main Effect</td>
<td>Non-Directional</td>
<td>Strangers and friends had the same average donation amount.</td>
<td>Podcast and talk show had the same average number of words recalled.</td>
<td>$IV_{B1} = IV_{B2}$</td>
</tr>
<tr>
<td>Interaction</td>
<td>Directional</td>
<td>There was a bigger difference from the little group to the big group for strangers than for friends.</td>
<td>There was a bigger difference from related words to different words for the podcast than for the talk show.</td>
<td>effect of $IVA$ for $IV_{B1} &gt;$ effect of $IVA$ for $IV_{B2}$</td>
</tr>
<tr>
<td>Interaction</td>
<td>Non-Directional</td>
<td>The effect of group size on the amount donated depends on the familiarity of group members.</td>
<td>The effect of word meaning on the number of words recalled depends on the type of distraction.</td>
<td>effect of $IVA$ depends on level of $IV_{B}$</td>
</tr>
</tbody>
</table>
5.3.2 Content Domains

Two different content domains were used at Time 1 and at Time 2. Time 1 graphs and statements referred to the amount donated (DV) based on group size (IVA) – little or big and group membership (IVB) – friends or strangers. Time 2 graphs and statements referred to word recall (DV) based on word meaning (IVA) – related words or different words and distraction type (IVB) – podcast or talk show. As appropriate, either the variable level or the category name was referred to in the statements. In this study, unlike in the studies presented in Chapters 4 and 6, participants did not have any additional background information about the source of the data represented in the graphs.

5.3.3 Graphs

Each graph contained two independent variables (IVA on the x-axis and IVB in the legend), each of which had two levels. The patterns specified in the simple effects consistency model (Figure 2.11) were used as the source of the data patterns. The values for the graphs were generated in the same way as in the Speeded Evaluation Task of Graph-Statement Matches (Chapter 6) as the details of the graph construction are more relevant to that task, I refer the interested reader to Section 6.2.3 for further details. A key feature is that the graph patterns were maximally different from one another given the constraints of the graph space. As there were 24 T/F judgments, some of the graphs were repeated. Eighteen of the twenty unique patterns were used;
5 of the graphs had 2 presentations each, 1 graph had 3 presentations. These variations in the frequency of each pattern helped balance the pattern of T/F responses.

At Time 1, there were not category labels for the two IVs; only the levels of the variables were labeled. Time 1 also did not have a numeric scale on the y-axis. Figure 5.2 shows the same graph pattern as it appeared at each Time.

**Stimuli at Time 1**

![Graph: Amount Donated]

**Stimuli at Time 2**

![Graph: Number of Words Recalled]

**Figure 5.2.** Examples of data patterns and graphs from Time 1 and Time 2. As shown here, Time 2 graphs had more of the typical formatting conventions represented (numeric scale for DV and category labels). The graphs for data pattern 3 are presented here.

### 5.3.4 Combining Graphs and Statements

Each statement was accompanied by 4 graphs. Graphs were assigned to statements based on the goal of having graph patterns from different sections of the theoretical graph space (see Figure 2.11) paired with each statement. In addition, having variation in T/F responses for each statement was also taken into consideration when assigning graphs to a statement.

The graph patterns as well as their order for a given statement were the same at both time points. However, the order of the statements was different at Time 1 and
Time 2, with the exception of the same hypothesized ‘easy’ statement (syntax: No effect of IV$_B$ at IV$_{A1}$) being first at both Times.

5.4 Procedure

The 6 statements, with 4 graphs each, were presented on paper. The subjects marked whether the statement was true or false for each graph. This resulted in a total of 24 responses per participant. The example problem presented earlier in Figure 5.1 shows one of the statements and its accompanying graphs, with the statement centered above the graphs. As shown in Figure 5.1, all graphs for a given statement were visible simultaneously.

Participants completed this task along with a graph construction problem (similar to those discussed in Chapter 4) and background information sheet. However, this does not imply that students completed the tasks in this order; two notes are relevant. Many students did not complete the construction problem. In addition, observation of the completion of the packets indicated that at least some students completed the graph-statement evaluation problems prior to the construction problem. Data on the construction problem is not considered.
5.5 Results

5.5.1 Exclusion of Subjects from Analyses

One of the statements participants evaluated was about the directional simple effect of IVB at IV\textsuperscript{A2} (Syntax: in IV\textsuperscript{A2}, IV\textsubscript{B1} > IV\textsubscript{B2}). Answering this question required participants to compare the magnitudes of two adjacent bars (see Figure 5.3). Across both timepoints, the majority (73%) of participants were at ceiling (correct on all 4 of the 4 graphs) on this question. Subjects who were wrong on at least 3 of the 4 graphs associated with this statement at either testing time were dropped from the analyses. Based on this criterion, 8 participants (7 lower math, 1 higher math) were dropped from the analyses. All but 2 of these 8 were excluded based on their performance at Time 2. Some of these dropped participants had previously answered all 4 graphs for this statement correctly at Time 1, but then responded correctly to only 1 of the same 4 graphs at Time 2. On other parts of the survey, they tended to also have considerable drops in performance. Together, this evidence suggests that these 8 subjects were not participating in the task. This leaves 37 subjects in the analyses; 89% of these subjects were at ceiling on the directional simple effect statement.

---

\textsuperscript{33} General Notes about Results: (1) Error bars on graphs represent one standard error unless otherwise noted. (2) For repeated measures ANOVAS in which the assumption of sphericity was violated, Huynh-Feldt corrections are used if epsilon > .75, else Greenhouse-Geisser corrections are applied.
Figure 5.3. Depiction of interpretation processes necessary for answering a directional simple effect statement. The statement used for exclusionary purposes, “With different words, subjects recalled more words when hearing a podcast than when hearing a talk show” was evaluated for four graphs, one of which is shown here. Most participants were at ceiling in identifying that the podcast magnitude was larger than the talk show magnitude. The steps shown in the figure comprise processes discussed in considering the levels of interpretation (Section 1.6.2).

5.5.2 Overall Performance

Overall, participants’ performance levels covered a considerable range at both timepoints. Subjects who participated at only one of the two times were not significantly different from their counterparts who participated at both times (independent samples t-tests, \( p > .4 \)). Figure 5.4 presents the individual scores, relative to the line for which Time 1 = Time 2 (i.e., \( y = x \)), on which subjects would fall if their performance was identical at Time 1 and at Time 2. Performance at the two time points was correlated, \( r(35) = .53, p = .001 \). On average, scores increased from Time 1 (\( M = .75, SE = .02 \)) to Time 2 (\( M = .79, SE = .02 \)), \( t(36) = 2.23, p = .03, Cohen’s d = .36. \)\(^{34, 35} \) As will

---

\(^{34}\) Treating missing problems as missing (such that accuracy for each statement was based on the number of graphs responded to rather than the four graphs presented with each statement), the change from Time 1 to Time 2 was not significant (\( p = .075 \)).

\(^{35}\) When the 8 excluded subjects were left in the analyses, there was not a significant different between Time 1 and Time 2.
be covered later (Section 5.5.3), this small amount of overall improvement arises from changes on a subset of the statements.

![Graph showing accuracy at Time 1 vs. Time 2](image)

**Figure 5.4** Average performance at Time 1 and Time 2, by subject. Participants’ overall accuracy levels varied widely at both timepoints, but significantly improved from Time 1 to Time 2. Scores above the line, at which Time 1 scores equal Time 2 scores, reflect improvement from Time 1 to Time 2 for an individual subject.

In total, 0.6% of the individual T/F responses to a unique graph-statement pairing were missing; scores for these pairings were treated as incorrect. This affected 4 participants at Time 1 and 1 participant at Time 2. Analyses were also conducted with
adjusted accuracies, in which subjects were scored based on the number of problems they completed. Results using adjusted accuracies are similar to those reported here.

Accuracy was above chance for 85% of the graphs when looking at accuracy levels across participants (20 of 24 problems at Time 1; 21 of 24 problems at Time 2). The binomial distribution was used to determine what accuracy levels would be different from expected by chance, using alpha = .05 and a two-tailed cut-off. In terms of a single graph-statement pair, if the average proportion correct was greater than approximately .53 or less than .29, performance the pair was significantly different from chance. Only a single graph-statement pairing was at or below the lower extreme of chance. This was on graph pattern 12 on the statement about the main effect of IV A; it was below chance at both time points (see Section 5.5.4 for further coverage of this).

Then, 3 additional pairings were at chance level. Combined with the pairings that were just above chance, these problems were primarily on the pre-test and dealt with the two statements about interactions.

All participants scored above .50, and for an individual subject, if the proportion correct was less than .30 or over .69, the participant’s overall performance was significantly different from chance responding. Figure 5.5 illustrates the pattern in which participants were more likely to improve than to decrease, relative to chance performance. A sign test based on whether participants increased or decreased their percent accurate indicated that significantly more increases were observed than expected by chance (23 of the 35 changes were increases, one-tailed \( p = .04 \)). (However, if considering only increases or decreases in which the total score differed by
at least 2 problems, the one-tailed sign test was only marginally significant, 14 of 20 changes were increases, \( p = .058 \).)

<table>
<thead>
<tr>
<th>Time 1</th>
<th>chance</th>
<th>above chance</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>21</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 5.5.** Counts of subjects’ standing relative to chance at Time 1 and Time 2. Scores between .30 and .70 were at chance, one-tailed binomial distribution with alpha = .05. Most subjects were above chance at both timepoints, and for those who started at chance, improvement (green shading on upper right) was the most frequent pattern.

Given the slight difference in average scores, knowing that more subjects increased than decreased is helpful, but it is also useful to know the size of those increases or decreases. The modal change was 1 problem for both increases and decreases. The average increase neared 3 problems, and the average decrease was 2 problems. **Figure 5.6** shows the changes scores (positive and negative) based on the 24 responses for each subject, relative to their overall performance level at Time 1. Participants scoring well to begin with had smaller changes. **Figure 5.6** presents an alternate representation of the performance levels presented in **Figure 5.4** by focusing on the change scores. Doing so indicates that additional variability in performance is present, as subjects tended to have performance changes in both directions.
 Increases from Time 1 to Time 2 (positive change scores) were more common. Participants performing well from the beginning had less change over all. Subjects tended to have both increases and decreases (i.e., some problems went from right to wrong while others went from wrong at Time 1 to right at Time 2).

Twelve participants were at chance at Time 1. In addition to being at ceiling on the directional simple effect statement used as an exclusion criteria, the relative accuracies levels of the different statements followed the general pattern of results from all subjects (see Section 5.5.3). This also applied to the 7 participants who were at chance at Time 2. Thus, given the similarity between their rankings of the statements and the ranking based on the sample as a whole, individuals who were at chance do not necessarily appear to have been performing randomly; rather their score seems reflective of their individual knowledge.

5.5.3 Effects of Statement Kind, Time, and Math

Subjects’ accuracy for each statement was assessed using a Time (2) x Statement (6) x Math (2) ANOVA with repeated measures on statement and time. The main effect
of time on accuracy was significant, $F(1, 35) = 4.22, p = .047, \eta_p^2 = .11$, as was the main
effect of statement, $F(3.8, 131.6) = 16.50, p < .001, \eta_p^2 = .32$, and a Time x Statement
interaction, $F(3.7, 131.0) = 3.42, p = .012, \eta_p^2 = .09$. Inspection of the pattern of means
(see Figure 5.7) suggests both the source of the time effect as well as the source of the
Time x Statement interaction. While scores on most of the statements were similar at
both time points, subjects significantly improved on the statements related to
interactions.

![Figure 5.7](image)

**Figure 5.7.** Accuracy based on statement at Time 1 and Time 2. Overall, collapsing
across time points, participants were more accurate on the two simple effect
statements than on the main effect and interaction statements, which did not differ
from one another. The two interaction statements significantly improved from Time 1
to Time 2. As discussed in the text, participants in the lower math group significantly
improved on the directional interaction statement ($p = .034$), and participants in the

---

$^{36}$ The effects for Statement and Time x Statement were also significant in the full
sample of $n = 45$. Performance improved on the non-directional interaction statement
and worsened on the directional simple effect statement, adjusted Bonferroni $ps < .02$. 
higher math group significantly improved on the non-directional interaction statement ($p = .018$).

Follow-up ANOVAs (Statement (6) x Math (2)) confirmed that the effect of statement persisted at both time points (Time 1: $F(3.2, 113.2) = 13.87, p < .001, \eta^2_p = .28$; Time 2: $F(4.1, 141.8) = 7.60, p < .001, \eta^2_p = .18$). At Time 1, the two simple effect statements did not differ from one another and were significantly higher than the two interaction statements as well as the statement about the main effect of IVB (adjusted Bonferroni $p$s < .01). Also at Time 1, the directional simple effect was evaluated more accurately than the main effect of IV_A, which in turn was evaluated more accurately than the main effect of IV_B (adjusted Bonferroni $p$s = .03). At Time 2, the two simple effect statements were still significantly easier than the IV_B main effect statement (adjusted Bonferroni $p$s < .05). Only the directional simple effect statement was significantly easier than the IV_A main effect statement and the directional interaction statement (adjusted Bonferroni $p$s < .01). The non-directional interaction statement did not differ from the other statements ($p$s > .1).

Overall, statements about simple effects remained the easiest at both time points. The accuracy of evaluating statements about the main effects depicted in the graph depended on which IV was referenced in the statement, although the main effect statement about IV_A was only significantly easier than the main effect statement about IV_B at Time 1. Evaluating interaction statements was the hardest, although the non-directional interaction statement was no longer significantly lower at Time 2.
Across subjects, a priori paired sample $t$-tests evaluating the effect of time at each statement indicated that improvement occurred only for the two interaction statements (directional interaction statement: $t(36) = 2.29, p = .028$, Cohen’s $d = 2.53$; non-directional interaction statement: $t(36) = 2.92, p = .006$; Cohen’s $d = 3.70$). On the directional interaction statement, the average proportion correct improved from $.61$ ($SE = .06$) to $.74$ ($SE = .05$). On the non-directional interaction statement, the average proportion correct improved from $.60$ ($SE = .05$) to $.80$ ($SE = .05$), at which point the non-directional interaction statement was not significantly different from the other statements. While these patterns occurred across both math groups, each group only had significant improvements on one of the statements. The lower math group significantly improved only on the directional interaction statement (Time 1: $M = .56, SE = .07$; Time 2: $M = .71, SE = .06$; $t(20) = 2.28, p = .034$). The higher math group improved only on the non-directional interaction statement (Time 1: $M = .72, SE = .26$; Time 2: $M = .91, SE = .18$), $t(15) = 2.67, p = .018$.

The analyses also yielded a main effect of math experience on accuracy (see Figure 5.8), such that participants with higher levels of math experience performed better ($M = .82, SE = .03$) than participants with less math experience ($M = .73, SE = .02$), $F(1, 35) = 6.47, p = .016, \eta^2_p = .16$. While there was not originally a Math x Time interaction ($p = .3$), the follow-up ANOVAs at the different time points revealed that

---

37 Using all participants who completed the task at Time 1 and Time 2 (rather than just those who met the inclusion criteria), those in the higher math group ($M = .81, SE = .03$) answered more accurately than those in the lower math group ($M = .70, SE = .02$), $F(1, 43) = 10.30, p = .003, \eta^2_p = .19$. The effect and effect sizes for math background are similar when using adjusted accuracy levels or when considering the original 45 subjects before individuals were excluded for difficulty on the directional simple effect question.
math experience had an effect only at Time 1, $F(1, 35) = 10.17, p = .003, \eta^2_p = .23$; Time 2: $p = .16$. As both math groups had room for improvement at Time 1 and Time 2, this does not appear to arise because of ceiling level performance. In addition, math experience did not interact with statement; that is, the pattern of difficulty across the statements was similar in both math groups (see Figure 5.8).38,39

38 The effects of Time on each statement were the same when analyses were run using adjusted accuracy scores in which participants’ accuracy levels were out of the number of problems completed.

39 Had the excluded subjects (Section 5.5.1) been retained, accuracy would have decreased from Time 1 to Time 2 on the statement about the simple effect of IV$_B$ at IV$_A$2. However, the overall ranking of the statements, effects of time, and effects of math would have been similar, although less likely to reach significance.
Figure 5.8. Effects of Math Level and Time on accuracy for each statement. In contrast to the previous figure, the performance is broken down by math group. Dashed lines represent the lower math group, while solid lines represent the higher math group. While the varying performance levels across statements can still be seen, the emphasis is on the differences between the math groups. Lower math participants had lower accuracies only at Time 1. Lower math participants significantly increased from Time 1 to Time 2 on the directional interaction statement, and higher math participants significantly increased from Time 1 to Time 2 on the non-directional interaction statement.
Analyzing the composite math variable which differentiated three skill levels led to a similar pattern of effects. The lowest math group was significantly worse than the highest math group, $F(2, 34) = 3.27, p = .05, \eta_p^2 = .16$, although again, this was only reliable at Time 1. The Time x Statement interaction, $F(3.7, 127.8) = 3.56, p = .01, \eta_p^2 = .10$, showed a similar pattern to what was found when math level was only based on math course. Only the interaction statements improved from Time 1 to Time 2. Differences among the statements at each time do pattern slightly different; the following results are based on pairwise comparisons of the statements using adjusted Bonferroni $p$ values < .05. For the lowest math group, it is the most similar. At Time 1, the above average math group only performed worse on the non-directional interaction statement relative to the directional simple effect statement. However, the average participants had no differences among statements at Time 2, and the above average group only performed better on the directional simple effect relative to the two main effect statements at Time 2. In general, the use of the composite background variable to establish 3 groups appears to separate out the very lowest and highest participants. However, as the average group only had 7 individuals (relative to 18 in the lower and 12 in the higher), the analyses focus on the two-way split based on math experience.

5.5.4 Effects of Graph Pattern

Performance levels also need to be considered within the context of the 4 graphs that participants evaluated for each statement. As the graphs were selected such that all graphs in the simple effects consistency model were used and such that there was variation in the correct responses, these are not systematic comparisons of the different
graph patterns across statements. On the other hand, the variation that arises does speak to how participants might be reasoning with the paired statement and graph. Logistic regressions modeling math (2) x graph (4) x time (2) with repeated measures on graph and time were run for each statement. In the following analyses, pairwise contrasts, using Bonferroni adjusted $p$ values, are used to compare the graph patterns when there was a main effect of graph on a given statement. These are referred to as graph-specific effects, as performance depended on the specific graph pattern that was presented for that statement. Results reported as significantly different are less than the Bonferroni-adjusted $p$ value of .05.

Each of the six statements will be considered in turn. The numbers used to refer to the individual graph patterns arise the graph patterns laid out in Figure 2.11 in the simple effects consistency model. The graphs in the model are numbered left to right, then top to bottom.

- There were no effects of graph pattern, math experience, or time on the directional simple effect statement. The ceiling performance on this question was expected given its use in setting inclusion criteria. Even so, however, the majority of participants in the full sample were already at ceiling on the question.
- The non-directional simple effect statement was only affected by math level, $F(1,35) = 5.20, p = .03$, with subjects in the higher math group ($M = .95, SE = .04$) being more accurate overall on this statement than subjects in the lower math group ($M = .83, SE = .04$).
- Performance on the non-directional main effect statement about IV$_A$ was also affected by graph pattern, $F(3, 105) = 26.65, p < .001$ (see Figure 5.10). In particular, for both time points and both math levels, performance on one of the graphs (Pattern 12) was significantly lower than all other graphs presented for this
statement: an average of .30 (SE = .06) compared to an average of above .90 (p < .001). In addition, Pattern 12 was also the only graph pattern in the whole task at the lower cut-off for chance. Re-examination of this particular graph pattern suggests that it might have been an issue of precision and/or perception. Among the remaining graphs presented with this statement, Pattern 10 was significantly higher than Pattern 20. While graph pattern did not interact with math level or time reliably, examining the means suggests that this difference between Patterns 10 and 20 was more pronounced in the lower math group.

- The non-directional main effect statement about IV was affected by graph pattern, \( F(3, 105) = 16.64, p < .001 \) (see Figure 5.11). Performance on graph pattern 16 was the highest. In addition, Pattern 9 was easier than pattern 17 and 18. Both math groups showed this pattern, although the trend was stronger in the lower math group.

- Performance on the directional interaction statement was affected by time, \( F(1,35) = 6.00, p = .02 \), and a math x graph x time interaction, \( F(3, 105) = 2.97, p = .04 \). The interaction appears to stem from lower math performing poorly on graph Pattern 4 at Time 1. In addition, it accounts for the main effect of Time, in that participants in the low math group improved on graph pattern 4 and participants in the high math group improved on graph pattern 9.

- The non-directional interaction statement was affected by math, time, and graph pattern. Accuracy was highest on graph Pattern 1, in which there were no main effects or interactions, than on the other graphs (patterns 8, 15, and 17), \( F(3, 105) = 3.10, p = .03 \). For this statement, there were also main effects of Time and Math. Time 2 (\( M = .81, SE = .05 \)) was higher than Time 1, (\( M = .62, SE = .05 \)), \( F(1,35) = 13.30, p < .001 \). The improvement from Time 1 to Time 2 primarily occurred on graph patterns g8 and g17 (see Figure 5.13). Higher math participants (\( M = .81, SE = .05 \)) also performed better than lower math participants (\( M = .61, SE = .05 \)), \( F(1, 35) = 6.67, p = .014 \).
Figure 5.9. Effect of math on the nondirectional simple effect statement. Accuracy about the lack of an effect of IV$_B$ at IV$_{A1}$ was relatively high across math groups and time points.
**Proportion Correct (SE) per Graph**

**Main Effect of IV_A:**

“Related words and different words had the same average number of words recalled.”

(Syntax: IV_A1 = IV_A2)

<table>
<thead>
<tr>
<th>Graph</th>
<th>Proportion Correct (SE)</th>
<th>Correctly Responded</th>
</tr>
</thead>
<tbody>
<tr>
<td>g3</td>
<td>.90 (.04)</td>
<td>TRUE (g3)</td>
</tr>
<tr>
<td>g10</td>
<td>.97 (.02)</td>
<td>FALSE (g10)</td>
</tr>
<tr>
<td>g20</td>
<td>.86 (.04)</td>
<td>TRUE (g20)</td>
</tr>
<tr>
<td>g12</td>
<td>.30 (.06)</td>
<td>TRUE (g12)</td>
</tr>
</tbody>
</table>

**Figure 5.10.** Effect of graph pattern on the nondirectional main effect statement for IV_A. From the means, this can be seen primarily from the low level of performance on graph 12. In addition, accuracy on graph 10 was significantly higher than on graph 20. These variations in performance did not interact with math or time.

*Note:* The statement and graphs shown are those used at Time 2, but accuracy levels reflect performance averaged across both timepoints.
Proportion Correct (SE) per Graph

Main Effect of IV\_B:
“Podcast and talk show had the same average number of words recalled.”

(Syntax: IV\_B1 = IV\_B2)

![Figure 5.11. Effect of graph pattern on the nondirectional main effect statement for IV\_B. Performance was higher on patterns 9 and 16, compared to patterns 17 and 18.](image)

Note: The statement and graphs shown are those used at Time 2, but accuracy levels reflect performance averaged across both timepoints.
Proportion Correct (+ SE) per Graph

Non-Directional Interaction Statement:
“The effect of word meaning on the number of words recalled depends on the type of distraction.”
(Syntax: the effect of IV_A depends on the level of IV_B)

Figure 5.12. Effect of graph pattern on the non-directional interaction statement. Accuracy was highest on graph pattern 1, in which no interaction was present.

Note: The statement and graphs shown are those used at Time 2, but accuracy levels reflect performance averaged across both timepoints.
Figure 5.13. Interaction between graph pattern and time on the nondirectional interaction statement. Gains from pre-test to post-test were also affected by the particular graph pattern on the nondirectional interaction statement; patterns 8 and 17 (at Levels 6,3 and 3,6 respectively) increased while graph 1 (with no interactions) had no change and graph 15 (at Levels 5,5) did not significantly increase. The increases are appearing on the graphs for which one main effect and an interaction were present.

5.6 Summary and Discussion

Overall, participants improved their accuracy of evaluating statements about interactions over the term of the undergraduate statistics courses in psychology. There was little change for other statement types. Their overall levels of performance at Time 2 still leave room for improvement, especially when interpreting statements about main effects and interactions. Accuracy for statements about the depiction of simple effects in graphs was at ceiling levels at both timepoints.

The accuracy of evaluating statements about the main effects depicted in the graph depended on which IV was referenced in the statement. Evaluating the main
effect for \( IV_A \) (the IV on the x-axis) was at an intermediate level of difficulty at both time points. Evaluating the main effect of \( IV_B \) (the IV in the legend) was hard at both time points. It is possible that this is an inherently more difficult task perceptually – finding the main effect for \( IV_B \) requires averaging non-adjacent bars whereas finding the main effect for \( IV_A \) requires averaging adjacent bars. The discrepancy in difficulty levels of main effects based on the IV fits with previous work by Carpenter and Shah (1998) in which interpretations of the graph tended to leave out the quantitative relationship between the legend variable and the DV (see section 1.6.1.2).

Evaluating interactions went from being among the hard statements at Time 1 to being at an intermediate level of difficulty at Time 2. This was predicted by a priori hypotheses as this was predicted to be the most difficult initially as well as a novel topic covered in the course. Interactions still remained difficult to work with.

While the interactions between performance and math level are important to note, it is also critical to remember that the pattern of effects was similar across the two levels of math experience. In addition, the groups were only significantly different at Time 1. This is important to note as it indicates that the lower math group was raised to the level of the higher math group.

The effect of graph pattern on some of the statements could have arisen from a bias to respond False. This could account for the graph specific effects for the main effect of \( IV_B \) (Figure 5.11), for the relative ease of Graph 10 for the main effect of \( IV_A \) (Figure 5.10), as well as the relative ease of Graph 1 for the nondirectional interaction
statement (Figure 5.9). However, as 67% of the correct responses were true, a bias to respond false would likely have impaired performance more overall.

Another possible explanation for the variation among the graph patterns for a specific statement is the number of averaging operations and comparisons needed to evaluate the content of the statement against the graph representation. The presence of graph-specific effects for several of the statements is an important avenue for further investigation. It may lead to a greater understanding of the strategies that participants are using to evaluate the statements. Possibilities under consideration include whether performance changes arise from precision issues when applying strategies, difficulty with the strategies, or misunderstanding of the conceptual relationship specified by the statement. For instance, in Figure 5.9, while a false bias could account for the ease of graph pattern 1, there could be a conceptual explanation. Perhaps it is easier to know that there is no interaction (clearly evident with graph pattern 1 in which there are no effects) than to decide what actually constitutes an interaction. Despite the lack of any variation in the graph pattern 1, participants were not at ceiling in stating that there was not an interaction.

Reinforcing this point is the various levels of performance present with the 6 graphs that each had two presentations. Rather than considering it as graph-specific effects for a particular statement, one can also consider the variation in the performance as statement-specific effects on a particular graph. For graph pattern 9, the proportion correct across the 3 statements ranged from .58 – .77 correct at Time 1.
For graph pattern 18, the two average performance levels were .40 and .96 at Time 1.\textsuperscript{40}

In the subsequent chapter, the graph-specific (and/or statement-specific) effects are investigated more systematically.

\textsuperscript{40} These accuracies include participants who only participated at Time 1.
CHAPTER 6

SPEEDED EVALUATION OF GRAPH-STATEMENT MATCHES

The finding in Chapter 5 that interpretation accuracy differed depending on both the statement and the particular graph indicates a need for further understanding of how the graph pattern and the statement jointly influence the application of reasoning processes. In particular, a more nuanced consideration of the different graph patterns and their components is necessary to further understand how graph pattern affects reasoning about interpretation statements. As discussed in Chapter 1 and Chapter 2, individuals are frequently presented with graphs that contain two independent variables (IVs). Part of interpreting these graphs is drawing conclusions about the relationships among the variables.

6.1 Overview

In the computerized task presented in this chapter, participants were presented with graphs of two independent variables and statements that described relationships among the variables. On each trial, participants saw a single graph and a single statement. The graphs were based on data with 2 IVs and 1 DV. The statements required participants to evaluate how the variables were related to one another in the graph. Undergraduates used the data presented in the graph to respond to the statement as quickly as possible while still being accurate. For ease of exposition, I start with an example (Figure 6.1) of a trial from the study and then move to an in depth
discussion of the stimuli and design. After discussing each of the materials, I return to how the task procedure uses them.

On average, subjects who viewed a low danger situation helped ________ subjects who viewed a high danger situation.
   a) The same amount as
   b) More often than
   c) Less often than
   d) [impossible to tell from the information given]

Figure 6.1. Example of a trial in the Speeded Graph-Statement Evaluation task. On each of 400 trials in this task, participants completed or evaluated a statement about the data patterns in the simultaneously presented graph. Here, an example of a trial is shown. In this example, a) the domain content of the graph is helping behavior, b) the type of interpretation statement is about the relationships between the averages of each level of IVB, and c), the graph pattern is pattern 18 (see section 2.5.3 for overview of graph patterns). The manipulation of each of these 3 parts of the task is discussed in the upcoming section.

Three aspects of this sample trial, shown in Figure 6.1, are worth noting: 1) the content of the graph, 2) the kind of statement, and 3) the data pattern shown in the graph. These three aspects were systematically manipulated over the course of this
experiment to assess how the data pattern and statement kind jointly influence graphical reasoning abilities.

6.2 Stimuli

6.2.1 Context Domains

In the trial shown in Figure 6.1, the graph and associated statement were about one DV (the frequency of helping) based on two IVs (the influence of perceived danger and bystander presence). This set of variables made up one of the four context domains used in this task. All of the contexts involved concepts that might appear in introductory psychology courses. The variables in the each of the contexts are listed in Table 6.1. However, participants were not given the structure and terminology (IVA, IVB, and DV) that is used to identify the variables listed in the table. Instead, prior to viewing the graphs, participants read several short paragraphs that introduced the context of the data collection and the variables referenced in the graphs and statements. The descriptions presented in these paragraphs are of the type that participants might plausibly encounter in basic psychology research or as case studies in statistics textbooks. Technical language, such as the terms independent or dependent variable, was avoided to not confuse participants with the terminology. The full text of all of the paragraphs is shown in Box 6.1.
Introductory Text Read by Participants (Distraction Context Domain):

In this study, researchers wanted to find out how word meaning and distraction type affected the number of words recalled. Subjects (20 per group) learned a list of words while listening to something before being asked to recall the words.

For word meaning, subjects learned a list of a) conceptually related words (e.g., bed, sleep, night ...) or b) conceptually different words (e.g., book, mouse, rainbow, ...). For distraction type, the list was learned while subjects heard a) podcast or b) a talk show.

Then, researchers measured the number of words recalled.

Introductory Text Read by Participants (Helping Context Domain):

In this study, researchers wanted to find out how perceived danger and bystander presence affected the frequency of helping. Subjects (25 per group) witnessed a staged argument that resulted in injury to one of the people before they had the opportunity to help.

For perceived danger, subjects saw an argument in which the type of injury indicated a) low danger (accidental) or b) high danger (intentional). For bystander presence, subjects were a) alone or b) accompanied.

Then, researchers measured the frequency of helping.

Introductory Text Read by Participants (Hand Grip Context Domain):

In this study, researchers wanted to find out how amount of sleep and sugar intake affected hand grip strength (kg). Subjects (30 per group) were asked to report what they usually did before gripping a special machine.

For amount of sleep, subjects indicated if they slept a) 6 hours per night or b) 10 hours per night. For sugar intake, subjects indicated if they ate a) 20 grams daily or b) 60 grams daily.

Then, researchers measured hand grip strength (kg).

Introductory Text Read by Participants (Emotions Context Domain):

In this study, researchers wanted to find out how job type and e-mail usage affected the number of emotions named. Subjects (23 per group) answered questions about themselves and their jobs before naming as many emotions as possible.

For job type, subjects indicated if their job involved a) working alone or b) working in a group. For e-mail usage, subjects indicated if they used e-mail a) rarely or b) constantly.

Then, researchers measured the number of emotions named.

Box 6.1. Introductory paragraphs for each of the four context domains. Each context had the same structure for its introductory text. The introductory text served to introduce the variables and their levels. Trials were blocked by context, and participants read about a context prior to the start of that block of the experiment.
The paragraphs introducing each context provided information about how the data in the graphs originated. Each context was based on a situation in which researchers wanted to find out how different ‘things’ affected a behavior. First, the goal of the data collection was introduced in terms of all of the variables, and the two independent variables were hypothesized to affect the dependent variable. Then, a short description was given of the task that generated the data. In turn, each independent variable and its two levels were listed. Then, at the end, the dependent variable was specified as what was measured. Thus, while the actual variables changed, the paragraphs presenting each context were structurally identical.

For three of the four contexts, the IVs could be interpreted as being on a continuous dimension (e.g., perceived danger could have intermediate values between low and high). However, the phrasing of the background paragraphs as well as the graph layouts indicated that the levels were to be considered discrete categories.

In contrast to the earlier studies, the background paragraphs used here identified the sample size of each group. This was done to inform participants that the groups were of equal sizes within each context, thus legitimizing unweighted averaging of the means for the IVs’ levels. Otherwise, the contexts are identical to those used previously.
Table 6.1. Independent and Dependent Variables used in the four context domains. $IV_A$ was the independent variable presented first in the background paragraphs, and $IV_B$ was presented second. In the graphs, $IV_A$ was always placed on the x-axis and $IV_B$ was in the legend. In the table, the two levels of the IV follow the IV name. The DV was always on the y-axis.

<table>
<thead>
<tr>
<th>Context Domain</th>
<th>$IV_A$ in the Context</th>
<th>$IV_B$ in the Context</th>
<th>DV in the Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>Helping Behavior</td>
<td>Perceived danger: low or high</td>
<td>Bystander presence: alone or accompanied</td>
<td>Percentage of Times Helped</td>
</tr>
<tr>
<td>Naming Emotions</td>
<td>Job type: working alone or working in a group</td>
<td>E-mail usage: rare or frequent</td>
<td>Number of Emotions Named</td>
</tr>
<tr>
<td>Hand Grip Strength</td>
<td>Amount of sleep: 6 hours per night or 10 hours per night</td>
<td>Sugar intake: 20 grams daily or 60 grams daily</td>
<td>Hand Grip Strength (kg)</td>
</tr>
<tr>
<td>Word Recall</td>
<td>Word meaning: related words or different words</td>
<td>Distraction type: podcast or talk show</td>
<td>Number of Words Recalled</td>
</tr>
</tbody>
</table>

A note on terminology. When graphs were presented for the subjects, the independent variable (IV) named first in the introductory paragraphs was always on the x-axis. In this text, this variable is $IV_A$ while the variable representing the groups in the legend is $IV_B$. In addition, when variables were referred to in the statements, the levels were always presented in the same order. Thus, taking the Helping Behavior context (see Figure 6.1), note that the order in which the variables and their levels are mentioned matches the order from the introductory paragraphs for the context (see Box 6.1). $IV_A$, perceived danger, has two levels: $IV_{A1}$ – low danger and $IV_{A2}$ – high danger. The second independent variable, $IV_B$, bystander presence, also has two levels:
The dependent variable (DV) is the percentage of times helped.

6.2.2 Statements

The example trial in Figure 6.1 presents 1 (“On average, subjects who viewed a low danger situation helped _________ subjects who viewed a high danger situation.”) of the 10 statements that participants responded to during the study. In particular, this statement dealt with one of the main effects displayed in the graph. Overall, the 10 statements used in the task referenced the main effects, simple effects, and interactions displayed in the graphs. The statements used in this task are of two kinds: multiple choice (MC) and true/false (T/F). The multiple choice statements presented response choices to what were essentially fill-in-the-blank questions. MC statements about the main effects and simple effects displayed in the graph had response choices that compared the frequency of the DV (Box 6.2, example 1). MC statements about the interactions displayed in the graph had response choices that compared the size of a variable’s effect (Box 6.2, example 2). T/F statements did not have missing words, and participants were given response choices relating to the validity of the statement (Box 6.2, example 3). The term ‘statement’ is used for both the MC and T/F versions because the participants’ completion of the MC question results in an interpretative statement about the data in the graph.

Table 6.2 presents the full wording of each statement (for the helping content domain). The phrasing of a particular statement was identical across context domains.
and as consistent as possible across the statement types. The wording of the statements was based on review of the linguistic literature on comparative statements as well as pre-testing of the comprehensibility of different syntax structures. The notation used to describe this phrasing is discussed in Box 6.2. Abstract, syntactic versions are used to focus on the concept of the comparison and to generalize across the different domains.

In the last column of Table 6.2, “Syntax of Statement”, the abstract labels of the independent variables and their levels are used to denote the comparison implicit in that statement. The question mark indicates that the study participant was asked to indicate the nature of the relationship between those two quantities; this is represented by the blanks left in the statement examples in Figure 6.3. Thus, $IVA_1 \ ? \ IVA_2$ indicates that the statement asked participants the nature of the relationships between the two levels of $IVA$. (Box 6.2 offers a further description of this terminology.)
Box 6.2. The notational scheme for the independent and dependent variables that participants read about during this study. These notations are used in this text to refer to abstract, syntactic versions of the statement content.

As mentioned earlier, participants were asked about main effects, simple effects, and interactions. Both the main effect statements and the interaction statements had non-directional and directional versions. Non-directional statements only asked if an effect was present; directional statements went a step further to ask about the direction of the relationship among the variable – that is, which level in the comparison had a greater value. The non-directional versions were presented in the T/F statements, whereas the directional versions occurred in the MC statements.

Statement 3 (“in IV\textsubscript{B2}, IV\textsubscript{A1} ? IV\textsubscript{A2}”) could have had the corresponding pair of “in IV\textsubscript{B1}, IV\textsubscript{A1} ? IV\textsubscript{A2}”, but pilot data (n = 9) indicated that participants responded with the same accuracy levels and similar RTs to these questions. This also applies to statement 4 (“in IV\textsubscript{A1}, IV\textsubscript{B1} ? IV\textsubscript{B2}”) and its corresponding pair. In addition, because the graph stimuli were constrained, the non-chosen pairs of statements 3 and 4 would have had little variability in the correct response. Thus, only the more variable member of each pair was used in this experiment.
addition, the statements occur in pairs (although presented in a random order). The 10 statements can be divided into five sets, with versions that focus on \( IVA \) and versions that focus on \( IVB \). In the interaction statements, both variables are involved, but one of IV is focused on as being the variable with an effect, whereas the other IV is referenced by its two levels.

**Box 6.3.** Sample statements from one of the content domains. The three selected statements illustrate the three different types of response choices that were presented. The syntactic form is indicated as assistance to the reader in recalling the role of the variables in the context; participants did not see this information.

*Note:* The unusual lettering choices assigned to the responses reflect the participants’ maintenance of consistent hand positions on the computer keyboard.
Table 6.2. Statement types presented as stimuli. Six statements were multiple choice (MC) and asked about the direction of the effect, with responses choices involving same, more, and less. Four statements were non-directional and true/false (T/F). In the MC statements, the “_____” indicates the missing words to be filled in with the appropriate comparative relationship between the two quantities indicated in the statement. In the syntax versions of the statements, a “?” is used instead to indicate that the relational term has to be selected by the participant. While the syntactic construction of the statements was replicated across contexts and statements, the response choices were tailored to the particular domain content and grammar of the statement.
<table>
<thead>
<tr>
<th>Statements (in Helping Context)</th>
<th>Response Choices (in Helping Context)</th>
<th>Effect Type</th>
<th>Syntax of Statement</th>
</tr>
</thead>
</table>
| On average, subjects who viewed a low danger situation helped _______ subjects who viewed a high danger situation.                                                                                                           | a) The same amount as  
as) More often than  
d) Less often than  
f) [impossible to tell from the information given]                                                                                              | main effect of IV<sub>A</sub> | IV<sub>A1</sub> ? IV<sub>A2</sub> |
| On average, subjects who were alone helped _______ subjects who were accompanied.                        |                                                                                                                                                                                                     | main effect of IV<sub>B</sub> | IV<sub>B1</sub> ? IV<sub>B2</sub> |
| When viewing accompanied, subjects who viewed a low danger situation helped _______ subjects who viewed a high danger situation.                                                                                           |                                                                                                                                                                                                     | simple effect of IV<sub>A</sub> | in IV<sub>B2</sub>, IV<sub>A1</sub> ? IV<sub>A2</sub> |
| When viewing a low danger situation, subjects who were alone helped _______ subjects who were accompanied.                                                                                                           |                                                                                                                                                                                                     | simple effect of IV<sub>B</sub> | in IV<sub>A1</sub>, IV<sub>B1</sub> ? IV<sub>B2</sub> |
| Danger level had _______ when viewing alone than viewing accompanied.                                    | a) The same effect on intervening  
s) A bigger effect on intervening  
d) A smaller effect on intervening  
f) [impossible to tell from the information given]                                                                 | Interaction (compare simple effects of IV<sub>A</sub>) | IV<sub>A</sub> had ? effect for IV<sub>B1</sub> than IV<sub>B2</sub> |
| Bystander presence had _______ when viewing a low danger situation than when viewing a high danger situation.                                                                                                           | Interaction (compare simple effects of IV<sub>B</sub>)                                                                                                                                               | IV<sub>B</sub> had ? effect for IV<sub>A1</sub> than IV<sub>A2</sub> |
| The percentage of times intervening depended on danger level.                                                                                                   |                                                                                                                                                                                                     | main effect of IV<sub>A</sub> | DV depended on IV<sub>A</sub> |
| The percentage of times intervening depended on bystander presence.                                                                                             | j) true  
k) false                                                                                                                                                                                 | main effect of IV<sub>B</sub> | DV depended on IV<sub>B</sub> |
| The effect of bystander presence was affected by the level of danger level.                                                                                     |                                                                                                                                                                                                     | interaction              | effect of IV<sub>A</sub> affected by level of IV<sub>B</sub> |
| The effect of danger level was affected by the level of bystander presence.                                                                                      |                                                                                                                                                                                                     | interaction              | effect of IV<sub>B</sub> affected by level of IV<sub>A</sub> |
6.2.3 Graph Stimuli

The data pattern shown in the sample trial in Figure 6.1 is one of 20 possible graph patterns. The 20 graph patterns conformed to the theoretical graph space based on simple effect consistency (see Section 3.3). With 20 graph patterns per domain and 4 domains, a total of 80 graphs were generated in Matlab for use in this experiment. Figure 6.2 shows one of the graph patterns and labels the quantities with the terminology and notation used in this text. Box 6.2 and Section 1.3.2 both give detailed descriptions of this notation.

Figure 6.2. Graph showing notation used to reference the quantities represented on the stimuli graphs. The notation is used to reference the quantities in the text; participants only saw the names of the variables and their levels. IV_A is the independent variable on the x-axis. IV_B is in the legend. The numbered subscripts indicate which level of the IV is being referenced; they go from left to right for IV_A and top to bottom for IV_B. Thus, in the sample graph shown here, IV_A1 is ‘6 hours’, IV_A2 is ‘10 hours’, IV_B1 is ‘20 Grams Daily’, IV_B2 is ‘60 Grams Daily’. Participants also read about the category labels for the IV themselves: IV_A is ‘Amount of Sleep’ and IV_B is ‘Sugar Intake’. The DV, ‘Hand Grip Strength (kg)’ is on the x-axis. Graphs were titled with the DV.
6.2.3.1 Constraints and procedures for generating data values

The graph stimuli presented to participants were bar graphs with two distinct levels per IV as shown in Figure 6.2. Thus, there were 4 magnitudes represented on each graph. One way of representing these values is in a 2 x 2 table. However, instead of focusing on the graph as displaying 4 separate values, it is helpful to use the slope-intercept form of linear functions to relate the points along a shared level of IVB. Thus, in order to clearly describe and show the set up of the graph patterns, line graphs will be used in this section. (This fits with Kosslyn’s (2006) and Cleveland’s (1994) recommendations that line graphs are better for conveying interaction patterns, even when the levels of the IVs are discrete.)

Thus, in generating the values for the 4 data points, the slope-intercept equation of the line is used for each level of IVB. This approach enables a clearer description and explanation of the manipulation of the relationship among the 4 data points. Thus, in the text that follows, the assignment of slopes and starting magnitudes to each level of IVB is covered. The starting magnitudes indicate the values for each level of IVB at IVA1. Then, when the slopes are combined with the starting magnitudes, the values for each level of IVB at IVA2 are obtained. These endpoints of the lines are the magnitudes plotted in the bar graph stimuli. Figure 6.3 identifies these components on the line graph version of the data plotted in Figure 6.2.
The starting magnitude of IV$_{B1}$ is determined by the axis shift.

The starting magnitude of IV$_{B2}$ is determined by the selected graph pattern.

The slope of IV$_{B1}$ is fixed at 38% of the vertical axis height.

The slope of IV$_{B2}$ is determined by the selected graph pattern.

Using $y = mx + b$ to generate the endpoints of the lines, under the assumption that the levels of IV$_a$ are located at the x values of 0 and 1.

**Figure 6.3.** Relation between slope-intercept form in a line graph and the bar graph stimuli. The slope-intercept form of a linear function can be used to more easily generate the data values to be used on the bar graph. The accompanying text walks through the assignment of the starting magnitudes of IV$_{A1}$ and IV$_{A2}$ as well as the slope of each line. These 4 quantities are indicated in this figure. Using the slope-intercept equation, $y = mx + b$, it is easiest to assume that the category of IV$_{A1}$ occurs at $X = 0$ and the category of IV$_{A2}$ occurs at $X = 1$. Using the intercept (and then the intercept + the slope) identifies the 4 endpoints, which are the magnitudes plotted in stimuli’s bar graphs.

Several constraints were built into this implementation of the simple effects consistency model to limit the number of possible graph patterns and reduce the influence of other aspects of the data pattern and graph representation. The graphs were constructed such that IV$_{B1}$ always had the same relative value for its slope.\(^{42}\) This

---

\(^{42}\) Even if one thinks about the stimuli in their actual bar format, the slope of the levels of IV$_{B}$ can still be referenced: it is the change in magnitude between IV$_{A1}$ and IV$_{A2}$ for the particular level of IV$_{B}$. 
was a constraint chosen to limit the number of possible graph patterns, while controlling for the influence of relative slope on the graph interpretation and perception. On all of the graphs, the slope of $IV_{B1}$ was just over $1/3$ (37.5%) of the height of the y-axis.

However, in order to lower the repetitive nature of the graph stimuli and to differentiate the data patterns that were presented in different context domains, two other processes affecting the starting magnitude of $IV_{B1}$ were used: 1) an additive axis shift and 2) a scaling factor. The axis shift varied for each graph, and the scaling factor varied across domains. The axis shift affected the proportional relationship between the starting magnitude of $IV_{B1}$ and the overall height of the y-axis. The scaling factor affected the numerical values of the starting magnitude of $IV_{B1}$.

The first process, the axis shift, varied quasi-randomly for each individual graph. This process amounted to selecting the relative location of the starting magnitude of $IV_{B1}$. A quasi-random constant was added to the data for each graph pattern. The variations were controlled such that the resulting graph pattern would be within the axis limits for that domain. Each value in a particular graph was raised by the same amount, thus retaining the relative nature of the relationships among the variables. This variation in the magnitude of the values relative to the overall height y-axis enabled the presentation of visually different magnitudes on each trial. As it was the same

---

43 Section 2.7 considered the other possible patterns that arise when this constraint is not imposed as well as how those graph patterns relate to the chosen set.
vertical shift for all 4 data values of a graph, the proportional relationships among the magnitudes remained the same.

Across domains, the scaling factor shifted, such that each domain was associated with a particular scale on the y-axis (maximum value on y-axis for the 4 domains: 160, 100, 60, and 20). The scaling factor was included to vary the numeric magnitudes on the axes and the values of the data points across domains. Thus, even though the same relative slope was used for IVB₁, the numeric values assigned to the data values would vary and would not always indicate a constant difference for the different domains. Together with the axis shift, the changes in the scaling factor enable the data values for the magnitudes represented on the graphs to take on different numerical values across within and across the different contexts.

Thus far, then, the starting magnitude and slope of IVB₁ are assigned. The axis shift determined the relative location of IVB₁ in regards to the axis. The scaling factor determined which values on the y-axis were associated with the representation of the data. The slope of IVB₁ was constant across graphs. The slope of IVB₂ was fixed relative to the slope of IVB₁, and it was based on the particular graph pattern.⁴⁴ The simple effects consistency model focuses on graphs that are maximally and equidistantly different from one another, across the whole graph space. This feature enables the assignment of slopes and starting magnitudes to IVB₂.

⁴⁴ Here, the technical aspects of the slopes are discussed. The conceptual relations between the different graph patterns can be found in Chapter 2.
6.2.3.2 Using the simple effects consistency model to generate data values

Stepping back for a moment to review, the 6 complexity levels proposed in the simple effects consistency model (see Section 3.3) are based on how consistent the sign and magnitude of an IV’s simple effects are to one another. The data patterns present at Levels 1 and 6 specifically define the slope of IV$_{B2}$. Figure 6.4 shows the graph patterns that result at these levels. That is, when both IVs are at Level 1, there are no main effects and no interactions and the data pattern must be two horizontal lines directly on top of one another. When both IVs are at Level 6, there are no main effects, but there is an interaction; thus, the data pattern must be lines of equal magnitude, but opposing slopes that cross at their midpoints.

![Figure 6.4](image)

Figure 6.4. Graph patterns are fully defined when both IVs are at Level 1 or Level 6. In the case of Level 1, there are no effects of any type, and the lines are coincident. In the case of Level 6, the main effects are cancelled out by the interaction, so the lines must have slopes of equal magnitude and opposite direction. They must cross at their midpoints.

When this model was introduced in Section 2.5.3, I focused first on the complexity level of each IV on its own before combining the two IVs. The same technique is helpful here. Starting with the IV on the x-axis, IV$_A$, the complexity level changes the degree of parallelism between the two lines.
As levels 1 and 6 have specific, pre-defined graphs, the intervening levels are separated from one another in equal units. Not only does this enable the use of graphs meeting each level’s definition – prototypical graphs for that particular level, but it also avoids the difficulty of assigning in between graphs to particular levels. (Section 2.7 discusses these assumptions.)

Thus, in going from one graph pattern to another, the variations that were possible in the simple effects were separated from one another in equal steps (see Figure 6.5). IV\textsubscript{B1} starts parallel to IV\textsubscript{B2} (Figure 6.5, level 1). Then, it ends (Figure 6.5, level 5) when IV\textsubscript{B2} is of equal, but opposing magnitude (i.e., directly opposite as if flipped over a horizontal line). In between these endpoints, you have the point at which IV\textsubscript{B2} is a horizontal line (Figure 6.5, level 3). Then, those sections are subdivided once again to have a line with half the slope of IV\textsubscript{B1} in either the positive (Figure 6.5, level 2) or negative (Figure 6.5, level 4) direction. This reflects the process that constructed each row of the theoretical graph space in the simple effects consistency model.

The starting magnitude of IV\textsubscript{B2} was also determined based on the graph pattern, as defined by the theoretical graph space. As you go from row to row in the graph space, there are systematic changes in the relative spatial location of IV\textsubscript{B2}. As shown in Figure 6.6, the starting magnitude of IV\textsubscript{B2} moves in equal units from below IV\textsubscript{B1} (Figure 6.6, row 3), to the midpoint of IV\textsubscript{B1} (Figure 6.6, row 6). The space in between is divided equally (Figure 6.6, rows 4 & 5).
Figure 6.5. Graphical representation of changes in IV_A’s complexity level. Establishing the slope of IV_B2 is based on the graph pattern. Returning to Figure 2.9, which is repeated here, the changes in slope from one level to the next occur in equally sized units. Starting with equal slopes (as shown by the coincident lines) at complexity levels 1 and 2, the slope of IV_B2 changes in equal steps (here, step size = 2) until it is equal in magnitude, but opposite in direction (at Level 6). The intermediate graphs are halfway between the extremes and the mid-point (level 4), for which the slope of IV_B2 is 0. (The depicted graphs at Levels 1 & 2 are at complexity level 1 for IV_B. The depicted graphs at Levels 3 – 6 are at complexity level 3 for IV_B.)

Note: Gridlines and numerical units for the slope are added for easier reference; they do not represent actual slopes used. Graph stimuli were bar graphs without gridlines. IV_B1 is a solid blue line, and IV_B2 is a dashed red line in this representation. IV_A1 is assumed to be at the left endpoints, with IV_A2 at the right endpoints of each line segment.
Figure 6.6. Graphical representation of changes in IV$_B$'s complexity level. IV$_{B2}$ systematically changes its relative spatial location going down a column (across the rows) of the theoretical graph space as its complexity increases. Returning to Figure 2.10, which is repeated here, the vertical translation of the line occurs in steps that are determined by the midpoint of IV$_{B1}$. In particular, in order to have equidistant positions, the step sizes are ¼ of the total distance (or ½ of the distance to the midpoint). (IV$_A$ is depicted at Level 4 in this figure.)
Note: Gridlines and numerical units for the slope are added for easier reference; they do not represent actual slopes used. Graph stimuli were bar graphs without gridlines. IV_B1 is a solid blue line, and IV_B2 is a dashed red line in this representation. IV_A1 is assumed to be at the left endpoints, with IV_A2 at the right endpoints of each line segment.

When these two processes are combined, they generate the 20 graph patterns described in the simple effects consistency model presented in Chapter 2. All 20 graphs used as stimuli in one of the contexts are shown in Figure 2.11.

Using these techniques to generate the data values for the graph stimuli makes the proportional relationships between the values on a given graph pattern the same across the domains. There are slight variations in magnitude due to the sizing of the graph based on the length of titles, but otherwise, the proportional relationships between the values are visually identical. The absolute magnitude of the values shifts based on the axis shift. The data values generated for each set of 20 graph patterns were tied to a particular context and not counterbalanced across domains or subjects.

6.3 Overall Design of Statement – Graph Pairings

The 20 graph patterns were crossed with the 10 statements, leading to 200 cells. In the design, each cell (i.e., each possible graph-statement pairing) was presented twice, for a total of 400 trials. The trials were distributed across the four domains based on Latin Square principles. Each of the 4 domains had 100 trials, and trials were blocked by domain. The blocks of each domain were presented in randomized order. Within each block, the trials were randomized.
An error in the design was fixed once noted; 38% of subjects completed the erroneous version of the task. The error resulted in an uneven distribution of the graph-statement pairings across domains because pairings were allocated incorrectly in the emotions domain. Thus, while each domain still had 100 trials, the subjects in the erroneous version of the design saw either 1 or 3 pairings of each possible graph-statement pair instead of uniformly seeing 2 pairings of each graph-statement pair. This applied to statements 2 through 5. In data analysis based on the 200 possible graph x statement pairings (a 20 x 10 design), only the first valid trial for each pairing was used. (Validity is defined in Section 6.6.)

**Table 6.3.** Correct responses based on statement type.

<table>
<thead>
<tr>
<th>Stmt #</th>
<th>Statement Type</th>
<th>Same As</th>
<th>More Than</th>
<th>Less Than</th>
<th>True</th>
<th>False</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>X1_avg ? X2_avg</td>
<td>0.3</td>
<td>0</td>
<td>0.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>G1_avg ? G2_avg</td>
<td>0.3</td>
<td>0.7</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>in G2, X1 ? X2</td>
<td>0.3</td>
<td>0.4</td>
<td>0.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>in X1, G1 ? G2</td>
<td>0.3</td>
<td>0.3</td>
<td>0.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>X had a ? effect for G1 than G2</td>
<td>0.2</td>
<td>0.8</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>G had a ? effect for X1 than X2</td>
<td>0.2</td>
<td>0</td>
<td>0.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>DV depended on X</td>
<td></td>
<td></td>
<td></td>
<td>0.9</td>
<td>0.1</td>
</tr>
<tr>
<td>8</td>
<td>DV depended on G</td>
<td></td>
<td></td>
<td></td>
<td>0.9</td>
<td>0.1</td>
</tr>
<tr>
<td>9</td>
<td>effect of X affected by level of G</td>
<td></td>
<td></td>
<td></td>
<td>0.8</td>
<td>0.2</td>
</tr>
<tr>
<td>10</td>
<td>effect of G affected by level of X</td>
<td></td>
<td></td>
<td></td>
<td>0.8</td>
<td>0.2</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>0.27</td>
<td>0.37</td>
<td>0.37</td>
<td>0.85</td>
<td>0.15</td>
</tr>
</tbody>
</table>

**Table 6.3** shows the frequency with which each possible response was correct. This is an uneven distribution because of how depicted data in the graphs relates to the types of statements participants interpreted. In particular, False was a correct response on only 15% of the trials. Alternate designs were considered, including adding
statements that had opposite response patterns (e.g., 90% false), but this would have made the task too long for participants. As will be discussed in the analyses, participants also responded differently to T/F problems than to MC, with more responses that were too fast or the wrong kind. Analyses of d prime values for the T/F questions are presented in Section 6.9.2.

In addition, as the correct answer was tied to the particular graph pattern, there is not a straightforward way to distinguish the effects of correct answer and graph pattern complexity.

6.4 Procedure

Like the example trial in Figure 6.1, each of the 400 trials presented a graph, a statement, and the appropriate response choices. The statement and response choices were presented in the upper left part of the screen (-150, 450 pixels from screen center), while the graph (840 x 630 pixels) was centered in the bottom part of the screen, graph center located at (0, -100) from screen center.

The experimenter talked participants through the instructions, the text of which was also presented on the screen (see Box 6.4). Participants were told that they would see two kinds of trials, and that they would be responding to the trials based on the information the graph conveyed. One kind of trial (the MC statements) had missing words; on those cases, participants were asked to choose the comparative statement to accurately complete the sentence. The other kind of trial (the T/F statements) did not have missing words; on those cases, participants were asked to respond true or false.
For both kinds of trials, participants had the option of indicating that they did not have enough information to tell how to respond to the statement. Participants were asked to respond as quickly as possible while still being accurate.

**Figure 6.7.** Schematic layout of task procedure. After receiving the instructions, participants viewed the introductory material for the first context. Then, there was a block of 100 trials for that context domain. This pattern, starting with the introductory material for the context, was repeated for each of the 4 contexts.

The response choices were visible on each trial, and participants typed in the letter that corresponded to their choice. The response choices were always presented to the participants in the same order. The choices were tailored to the particular

---

Across all participants, the response of “not enough information to tell” was used for 6.7% of the responses. Of these responses, 19% had RT less than 2 seconds. The frequency for each graph pattern averaged 5% (SD = 2%) of the ‘cannot tell’ responses. Hereafter, it is treated no differently than other wrong answers.
domain content (e.g., grammar based variations such as “fewer words than”, “less often than”), but the phrasing and terminology was matched as closely as possible. Upon typing in the letter for their response, the screen immediately cleared and proceeded to the next trial. Within the task itself, the program offered the opportunity for pauses every 20 trials.
Box 6.4. Instructions for the Speeded Graph-Statement Evaluation task. Participants viewed these written instructions on the computer screen while the experimenter talked through the instructions for the task.
As shown in the schematic in Figure 6.7, the trials were blocked by content domain, and the order of the content domains was counterbalanced across participants. Each content domain block was preceded by an informational screen which displayed the background paragraphs and sample graph layout. The background paragraphs (as discussed in Section 6.2.1) offered a brief description of the variables and the experiment leading to the data that was graphed. The sample graph layout (see Figure 6.8) had no data; it illustrated how the variables would be located on the graph. This information was intended to familiarize the participant with the variable labels and their spatial placement on the graph prior to the participant starting the timed trials.

![Figure 6.8](image)

Figure 6.8. Example of the graph layout participants saw as a sample. This was presented to participants at the start of each content block. All the graphs in the following content block followed this placement of variables.

6.4.1 Timing

Timing for each trial began from the initial (simultaneous) appearance of the graph, statement, and response choices. Timing for a trial ended when the participant
entered an appropriate response. In this context, appropriate responses are those which correspond to one of the possible response letters for any of the statement kinds (for MC: a,s,d,f and for T/F: j,k,f). However, if the participant responded with a different key (e.g., “m”), the computer program required the participant to re-enter their response until an appropriate letter was entered. Participants were informed during the instructions that the program would not move on to the next trial until an appropriate response was entered. There was not an additional prompt on the screen when re-entry was required. These are referred to as self-corrected (SC) trials as the participant was able to correct their response before moving onto the next trial.

6.5 Analyses Plan Based on Experiment Design

The overall structure of the experimental task resembles a Graph (20) x Statement (10) x Context Domain (4) x Trial (2) design with repeated measures on all of the factors. Reaction Time (RT) is available along with the participants’ accuracy based on their response to each trial.

Looking first at the 20 different graph patterns, the simple effects consistency model gives them a systematic order. The analyses make use of that order by focusing on the rows (complexity level of IV_b) and columns (complexity level of IV_a) that define the graph space. While there are 6 complexity levels at each IV, these levels cannot be fully crossed with one another. As a reminder, graph complexity levels 1-2 cross with one another as do levels 3-6. These two spaces (4 graphs, 16 graphs respectively) will be analyzed separately due to the non-continuous nature of the levels and different
dimension sizes. Levels 1-2 represent the graphs that do not have interactions, whereas Levels 3-6 represent graphs with interactions. Thus, the analyses are run separately for levels 1 and 2 and for levels 3 to 6. This removes the factor of graph (20) from the analyses and replaces it with either the factors \[(\text{complexity}_{IVA}(4) \times \text{complexity}_{IVB}(4))\] or the factors \[(\text{complexity}_{IVA}(2) \times \text{complexity}_{IVB}(2))\].

Returning to the overall design, the unique combination of a single graph with a single statement is referred to as a graph-statement pair. The graph-statement pairs categorize the factors that are of interest, and as such, are referred to as the cells for which subjects supplied data. Trials refer to the individual responses that subjects made. The cells collapse across the different domains and uniquely identify the possible graph-statement pairings.

However, as was mentioned in Section 6.3 and as will be discussed in Section 6.6, not all subjects had two valid responses for each possible graph-statement pair. On a by-subject basis, the first valid response for each cell was included in the analyses. This does discard more of the data, but it compares subjects on an equal footing.\(^{46}\) In the case of missing cells (for which none of the subject’s trials were valid), missing data was filled in (see Section 6.7.3) if a subject had 5 or fewer missing cells. Then, participants’ data was analyzed for the first valid cell per subject using a Statement (10)

\(^{46}\) Three percent of responses \((n = 151\text{ of }5800)\) for the first valid trial came from the subjects’ 2\(^{nd}\) trial. In addition, considering cells that were already missing and cells that would become missing if needing a second trial, 24% of the two-trial cells would have been missing with this subset of 29 subjects. Thus, analyses were limited to the first trial.
x IV_A_Complexity (2 or 4) x IV_B_Complexity (2 or 4) x Stats_Experience (2) with the mean-centered math background composite as a covariate.

Data from the full sample, based on a variable number of trials and cells per subjects, is reported in Appendix 1. These are included as it is relevant to compare the results from the analyzed subgroup of participants to the overall general pattern across all participants in the sample.

6.5.1 Summary of Design of Analyses

For each of the sets of analyses in Sections 6.8 – 6.12, the primary interest is in the effects of statement and the graph pattern (as represented by the complexity levels of IV_A and IV_B). In addition, a key question is the extent to which performance on the statement kind interacts with the graph pattern. Additional background variables of math and statistics experience are included in the design as well. These analyses are followed by consideration of two types of consistency: consistency of the subjects in their responses to the repeated trials of the same graph-statement pair, and b) consistency of the different individuals with the overall group averages.

Sections 6.8 and 6.9 report the results of ANOVAS that look at reaction time (Section 6.8) and accuracy (Section 6.9). As mentioned earlier, results from levels 1 and 2 are analyzed separately from results from levels 3 to 6 at this point for both statistical and conceptual reasons. The analyses are divided initially into the separate main effects associated with the statement and the complexity levels of the graph patterns. When the complexity levels of IV_A and IV_B had significant interactions with one another, the
Overall, two dependent measures (Accuracy and RT) are analyzed in this study. Within each section, the analyses of each dependent measure are done independently for each subset of graph patterns, based on the complexity levels from the consistency of simple effects model.

6.6 Participants

Fifty-three participants (19 M, 34 F) ran in the speeded evaluation of graph-statement matches task. Forty percent of the participants were in their first or second undergraduate year ($M = 21.0$ years, $SD = 5.3$ years). The students were undergraduates at a socioeconomically and ethnically diverse large northeastern university. All participants signed informed consent forms and received 2 movie tickets for their participation.

In order to be included in the analyses, participants had to have at least one valid response for every possible graph-statement pairing. (Validity is defined and discussed in detail in Section 6.7 – for now, it can be taken to mean that the subject was engaged and appropriately participating in the experiment.) In total, 29 participants (21 F, 8 M) were included in the analyses. For this group, 45% had a math/science major, 60% were in the high ($\geq 600$) math SAT group, and 66% were at the higher level of math experience (at least Calculus I). These three measures of math experience were
combined into an average math experience score, that was mean centered at 0. Statistics experience was also considered relevant to performance on this task; 55% of the participants had taken a statistics course.

6.7 Trial Validity and Missing Data

A general definition of valid trials would be that the subject was appropriately participating in the task. Both the participant’s response and his or her reaction time have an impact on whether a trial is valid or not. In brief, a participant had to spend a ‘reasonable’ amount of time to answer the question and had to select a response that was actually one of the listed response options. Table 6.4 presents the decisions about when to include a trial or not, and Table 6.5 presents the frequency of each type of removed trial.

Table 6.4. Characteristics of trials Included in analyses.

<table>
<thead>
<tr>
<th>Trial Included in Analyses?</th>
<th>RT &lt; 2 seconds (too fast)</th>
<th>RT &gt; 60 seconds (too slow)</th>
<th>Self-Correct (SC)</th>
<th>Wrong Kind (WK)</th>
<th>2 sec &lt; RT &lt; 60 seconds &amp; Valid Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>YES</td>
</tr>
</tbody>
</table>
Table 6.5. Rates of problematic trials in the data set.

<table>
<thead>
<tr>
<th>Rate of Trials in Experimental Data Set</th>
<th>RT &lt; 2 seconds (too fast)</th>
<th>RT &gt; 60 seconds (too slow)</th>
<th>Self-Correct (SC)</th>
<th>Wrong Kind (WK)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rate of Trials in Experimental Data Set</td>
<td>12% of the trials (range by subject: 0% to 49%, SD = 14%)</td>
<td>&lt; 1%</td>
<td>&lt; 1%</td>
<td>5% of the trials (range: 2% to 35%, SD = 7%)</td>
</tr>
</tbody>
</table>

6.7.1 Invalid Trials based on RT

On the lower end of possible RTs, a minimum RT of 2000 msec was required for the inclusion of a trial. This is a conservative requirement in that it is likely to miss other fast responses that are too fast to indicate actual processing.\[47\] Several participants showed clear patterns for their RT distributions in which responses that were made too early could be distinguished from responses that were not erroneous key presses; there was a peak in frequency of responses around 1-2 seconds before gradually starting to increase. The presence of trials that were too quick also became clear when looking at graphs of the cumulative RT.

Overall, subjects averaged 48 (SD = 58) (out of 400 total) fast trials per subject; 40% of the subjects had less than 10 trials with too fast RT. The large sd arises from

\[47\] However, if it is assumed that responses on trials that are too fast (e.g., < 2000 msec) are randomly distributed across the possible response choices, then the conservative approach to removing RTs that are too fast would not be expected to affect the pattern of findings. This assumption is empirically supported by the similar distribution of correct and wrong responses on the trials answered in less than 2 seconds.

Other approaches to reducing outliers are worth being considered (e.g., individual standard deviation – but this by itself is not effective because of either high rates of responding too fast or extreme outliers on the long end).
subjects who approached the cut-off point for being excluded for too many fast RT. However, the percentage of too fast trials differed by the type of statement. MC statements ($M = 10\%, SD < 1\%$) had fewer RTs that were too fast than T/F statements ($M = 15\%, SD < 1\%$).

At the upper end of possible RTs, RTs were required to be less than or equal to 60 seconds for inclusion of the RT in analyses. From the individual RT distribution curves constructed in Matlab for correct and incorrect responses, participants tended to have only infrequent responses above 45 seconds. Statement 5 possibly had a slightly higher probability of having long RT (2\% versus $\leq 1\%$) of all trials (across subjects and across graphs). While it would be possible to analyze responses for these trials as response accuracy and RT are analyzed separately, only responses that had valid, non-missing RTs were included in the analyses. (This eliminated less than 1\% of the possible trials).
Figure 6.9. Subjects’ cumulative RT and cumulative number of trials. Data for two subjects are presented. The first subject (332) remained on task throughout the experiment. The second subject (331) started responding very quickly to trials after the 1st 200 trials. This is seen in the leveling off of the slope of the lines. In addition, around the same trial, one also starts to see an increase in the slope of the number of incorrect responses, such that that rate becomes equal to that of the rate of correct responses.

Note. Green lines represent trials with correct responses. Dashed red lines are incorrect responses, and the thinner yellow line is responses of the wrong kind (see 6.7.2). The black line represents all trials.
6.7.2 Invalid Trials Based on Response

As indicated in the procedure (Section 6.4.1), the program required participants to enter an appropriate response (i.e., a response choice that was used at some point during the experiment). However, unfortunately this does not guarantee that the response was valid – that is, that the participant’s response was an actual response choice for that particular statement. Thus, all trials with invalid responses (e.g., responding with the key for true to a statement that was multiple choice) are discarded from all analyses; these are responses of the wrong kind.

Responses of the wrong kind were more frequent on the T/F statements (\(M = 14\%, \ SD = 4\%\)) than on the MC statements (\(M = 1\%, \ SD < 1\%\)). Part of this might be expected given the higher frequency of MC statements overall (6:4 ratio). However, as the T/F statements also had higher rates of responding too fast, it is possible that these differences in the rate of valid trials reflect a subject bias.

The remaining trials that were discarded from the data set were trials in which the participant self-corrected their response. The participant initially entered a response that was not a choice on any of the questions in the experiment. (This eliminated less than 1% of the possible trials).

6.7.3 Excluding Subjects and Working with Missing Data

All trials with too fast or slow of an RT as well as responses of the wrong kind or self-corrected responses are discarded from all analyses. Table 6.5 shows the relative
rates of each of these components. At this point, the decision was made about whether subjects had enough valid trials to be included in analyses.

The first criterion for inclusion was that over 60% of the 400 trials had valid data. Six subjects were discarded based on this requirement. Second, subjects had to have 5 or fewer cells for which there were no valid trials, out of the 200 possible cells (20 graphs x 10 statements). An additional 18 subjects did not meet this criterion and were excluded from these analyses. Of the remaining 29 subjects – the data set on which the analyses in this chapter are based – 12 of them had 5 or fewer missing cells (a total of 33 missing cells, 1% of the data), and values for those cells were calculated based on that subject’s responses.

The missing cells were filled in with that subject’s average accuracy level for that statement on the graphs that shared crossed levels (e.g., a missing cell for a graph on levels 1 or 2 was only based on a statement’s average for graphs 1 and 2 for that subject). This approach was selected as one of the key questions in the study is ‘given the statement, how well do subjects understand the data represented in a particular graph?’ With the statement being treated as a given, it is reasonable to fill in the missing cells with the average for that statement. In addition, the averaging was limited to the shared levels because of expected differences in graph pattern.
To briefly review, the analyses presented in this chapter are based on the first valid trial for each graph-statement pair for 29 subjects.

Across subjects, statements, and graphs, mean reaction time was 14.2 seconds (\(SE = .13\) sec, range: 9.1 to 21.8 seconds) (see Figure 6.10). The average proportion of correct responses ranged from .39 to .85 correct across subjects, for all of the statements and graphs (\(M = .70, SE < .01\)). The frequency of the 'cannot tell' response (that they could not determine the answer from the given information) ranged from 0% to 13% (\(M = 3\%, SD = 3\%)\).

Individual correlations between RT and accuracy for each subject were not significant, suggesting that there was not a trade-off between speed and accuracy. The individual correlations ranged from -.25 to .18 (\(M\) of the individual \(rs = -.05, SD = .10\)), \(p > .05\).

---

48 General Notes about Results: (1) Error bars on graphs represent one standard error unless otherwise noted. (2) For repeated measures ANOVAS in which the assumption of sphericity was violated, Huynh-Feldt corrections are used if epsilon > .75, else Greenhouse-Geisser corrections are applied.
Figure 6.10. Mean reaction time and accuracy levels by participant. The correlation between reaction time and average accuracy was not significant, for individuals or at a group level, $p > .1$.

Average reaction time, across trials, was higher for incorrect responses by 574 msec. Around 80% of subjects spent longer on problems that they answered incorrectly, as seen in Figure 6.11. Negative differences indicate longer reaction times for incorrect trials. Spending longer on wrong problems was significantly correlated with higher average accuracy scores, $r(27) = -.58$, $p = .001$. 
Figure 6.11. Reaction time differences based on problem accuracy. Longer reaction
times on incorrect problems were significantly correlated with higher accuracy levels,
\( r(27) = -.58, p = .001 \). Negative difference scores indicate participants spent longer on
incorrect problems. Each point represents the difference between the average RTs for
correct and incorrect trials for each participant.

6.9 Results: Effects of Statement

6.9.1 Effects of Statement on RT

Working with the graphs at complexity levels 1 and 2, the repeated measures
ANOVA on reaction time yielded a main effect of statement, \( F(8.2, 150.8) = 2.89, p =
.004, \eta^2_p = .10 \) (see Figures 6.12 and 6.13). However, pairwise comparisons (\( p \) values
Bonferroni adjusted) did not indicate any significant differences among the individual
statements, although the directional (MC) interaction statement focusing on IV_A took
marginally less time than the IV_B simple effect (MC) statement (\( p = .07 \)).

There was also a main effect of statement for graphs at complexity levels 3 to 6,
\( F(7.3, 189.7) = 23.18, p < .001, \eta^2_p = .47 \) (see Figures 6.12 and 6.13). Out of all the
statements, participants responded most quickly to the T/F statements about the main
effects. From Figures 6.12 and 6.13, it is apparent that asking about IV_A was no different from asking about IV_B, for main effects, simple effects, or interactions.

Pairwise comparisons on the main effect of Statement showed no differences among the IV_A and IV_B versions of the statements (adjusted Bonferroni p values > .05). In addition, however, it is also clear that participants spent more time on MC questions than on T/F versions of the same content. Pairwise comparisons on Statement indicated that this was true for the main effect statements (adjusted ps < .01) as well as for the interaction statements (adjusted ps < .006).

Given the difference between MC and T/F response times, the focus will be on the pairwise comparisons across the statement kinds, but within the same question modality. On the T/F statements, only the IV_B main effect statement was significantly faster than the second interaction statement (stmt 10) (adjusted p = .047). The other pairwise comparisons were not significant. On the MC statements, the two interaction statements took significantly longer than all other MC statements except the IV_A main effect statement. In addition, the two MC simple effect statements took significantly less time than the IV_A main effect statement. Evaluating the simple effect of IV_A was not significantly different than evaluating the two main effects, but evaluating the simple effect of IV_B took significantly less time than evaluating the main effects. However, interpreting these differences needs to be undertaken cautiously as even though the question modality was constant for the comparisons discussed above, there was variation in the syntactic structure and length of the statements.
Differences in statement length and syntactic complexity preclude in depth comparison of RTs across different statement types. As the question length and syntactic complexity varied by statement, the comparisons of RTs across statements are limited to those pairs which had equivalent syntactic requirements (e.g., the two MC statements about main effects; the two T/F statements about main effects, etc.). This lets the focus be on the content implied by the statement and not on other contributing factors from the syntax. (However, the impact of syntactic factors was likely reduced as each syntax is seen in 80 trials over the course of the experiment.)

Comparing the reaction times for graphs without interactions (Levels 1-2) to reaction times for graphs with interactions (Levels 3-6) (representing RT for complexity levels 3-6), it is clear that participants responded faster overall to graphs that did not have interactions. However, it is also clear that this depended on the statement type. The difference in reaction time primarily occurred on directional (MC) statements about main effects and directional statements about interactions; for those cases, graphs that did not depict interactions were faster.
Figure 6.12. Reaction times by statement: Simple effects and main effects. Performance varied by statement type within a given set of graphs (with or without interactions). In addition, statements regarding graphs without interactions were responded to more quickly for directional (MC) main effect statements.
Figure 6.13. Reaction times by statement: Interactions. Statement type had a larger influence on reaction times for graphs that were at Levels 3-6 than for graphs at Levels 1-2. The directional (MC) statements about interactions took longer than the non-directional (T/F) statements about interactions only for graphs that depicted interactions.

6.9.2 Effects of Statement on Accuracy

For the 80 T/F responses for each participant, d prime was calculated for each subject. These values are plotted in Figure 6.14. Values for d’ were positively correlated with accuracy on the T/F statements, $r(27) = .62, p < .001$. Subjects with higher d’ values had higher levels of accuracy on those statements.
**Figure 6.14.** Values of d prime for T/F statements. The majority of participants had d’ values between 0.5 and 2.

For the graphs without interactions (complexity levels 1 and 2), there was a main effect of Statement, $F(9, 243) = 6.04, p < .001$. The two T/F statements about interactions were lower than most other statements. For the graphs with interactions (complexity levels 3 to 6), there was a main effect of Statement, $F(9, 243) = 36.46, p < .001$. Table 6.6 reports the mean accuracy levels for each statement for graphs with and without interactions. Overall, it is clear non-directional interaction statements were always hard. The directional interaction statements were hard when the graphs had interactions, but they were not as difficult when the graphs did not have interactions. The opposite switch occurred with main effects. Statements about directional main effects were at intermediate levels (to easy on Levels 1-2) for both levels, but performance on the non-directional main effect statements was hard for graphs with interactions. Simple effects remained at intermediate – easy levels of performance. When there were differences (e.g., non-directional main effect statements), the
differences were more likely to be in the direction that the statement about IVB was harder, although both versions occurred.

**Table 6.6.** Mean accuracy levels by statement and graph pattern set.

<table>
<thead>
<tr>
<th>Levels 1-2</th>
<th>Type of Statement</th>
<th>Statement Focused on IV&lt;sub&gt;A&lt;/sub&gt;</th>
<th>Statement Focused on IV&lt;sub&gt;B&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Simple Effects</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Directional (MC)</td>
<td>0.86</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>Directional (MC)</td>
<td>0.84</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>Main Effects</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Non-directional (T/F)</td>
<td>0.81</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>Directional (MC)</td>
<td>0.79</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>Interactions</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Non-directional (T/F)</td>
<td>0.62</td>
<td>0.64</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Levels 3-6</th>
<th>Type of Statement</th>
<th>Statement Focused on IV&lt;sub&gt;A&lt;/sub&gt;</th>
<th>Statement Focused on IV&lt;sub&gt;B&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Simple Effects</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Directional (MC)</td>
<td>0.83</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>Directional (MC)</td>
<td>0.77</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>Main Effects</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Non-directional (T/F)</td>
<td>0.55</td>
<td>0.49</td>
</tr>
<tr>
<td></td>
<td>Directional (MC)</td>
<td>0.59</td>
<td>0.53</td>
</tr>
<tr>
<td></td>
<td>Interactions</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Non-directional (T/F)</td>
<td>0.68</td>
<td>0.59</td>
</tr>
</tbody>
</table>
6.10 Results: Effects of Graph Pattern

6.10.1 Effects of Graph Pattern on RT

For the graphs without interactions (complexity levels 1 and 2), there was a main effect of IVB Complexity, \( F(1, 26) = 14.31, p = .001, \eta^2_p = .36 \) and ofIVA complexity, \( F(1, 26) = 13.89, p = .001, \eta^2_p = .35 \). The reaction times for these pairs are shown in Table 6.6. For both IV\(_A\) and IV\(_B\), participants took longer on graphs at Level 2 than at Level 1.

Analyzing the RT for the graphs at the higher complexity levels, there was a main effect of the complexity level of IV\(_B\), \( F(3, 78) = 3.73, p = .02, \eta^2_p = .13 \), and a main effect of the complexity level of IV\(_A\), \( F(3, 78) = 3.45, p = .02, \eta^2_p = .12 \). Table 6.7 shows the influence of the complexity of each variable on reaction time. Pairwise comparisons for the main effect of the complexity level of IV\(_A\) indicated that participants took significantly longer responding to complexity level 5, than to complexity level 4 (adjusted Bonferroni \( p = .007 \)). Although complexity level 3 was also lower than complexity level 5, this difference was only marginally significant (\( p = .06 \)). None of the pairwise comparisons for the main effect of the complexity level of IV\(_B\) were significant, despite the values being similar to those for IV\(_A\). (While RT means were not statistically compared across the two sets of complexities, it is apparent that RTs were faster for graphs that did not depict interactions.)
Table 6.7. Reaction time based on graph pattern (across all statements)

Reaction Time Across Statements
$M$ in sec ($SE$)

<table>
<thead>
<tr>
<th>IV$_A$ Complexity</th>
<th>IV$_B$ Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1 10.9 (0.4)</td>
<td>Level 1 10.7 (0.6)</td>
</tr>
<tr>
<td>Level 2 12.5 (0.6)</td>
<td>Level 2 12.7 (0.6)</td>
</tr>
<tr>
<td>Level 3 14.5 (0.6)</td>
<td>Level 3 14.6 (0.5)</td>
</tr>
<tr>
<td>Level 4 14.5 (0.7)</td>
<td>Level 4 14.4 (0.6)</td>
</tr>
<tr>
<td>Level 5 15.4 (0.6)</td>
<td>Level 5 15.5 (0.6)</td>
</tr>
<tr>
<td>Level 6 15.1 (0.5)</td>
<td>Level 6 15.0 (0.6)</td>
</tr>
</tbody>
</table>

Note: Mean reaction times for each complexity level are given; standard error is in parentheses. For both IVs, Level 2 took significantly longer than Level 1. For IV$_A$, Level 5 took significantly longer than Level 4 ($p = .007$) and marginally longer than Level 3 ($p = .06$). The values are collapsed across the other dimension of the graph space. These RT values are estimated at the mean-centered Math Composite level.

6.10.2 Effects of Graph Pattern on Accuracy

As seen in Table 6.8, there were main effects of both IV$_A$ Complexity and IV$_B$ Complexity on accuracy for both sets of graph patterns. The means are reported in Table 6.9. Simple effect analyses and pairwise contrasts, $p < .05$, were used to determine which levels were significantly different from one another. Overall, for each IV, participants were more accurate for graphs at Level 1 than for graphs at Level 2. For Levels 3 to 6, the general trend is the highest accuracy for graphs at Level 3 and the lowest accuracy for graphs at Level 6. For both IV$_A$ and IV$_B$, accuracy was lowest on Level 6. For IV$_A$, complexity level 3 was marginally easier than levels 4 and 5, which did not differ from one another. For IV$_B$, complexity level 3 was significantly easier than levels 4 and 5, which again did not differ from one another. Thus, the pattern appears
to be that the endpoints (levels 3 and level 6) are significantly easier or harder than the graphs at the intermediate levels. However, as can be noted from the nearly identical means for both IVs, the overall trends are similar.

Table 6.8. Statistical results for the main effects of complexity for each IV.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Graph Pattern in Analysis</th>
<th>Inferential Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>IV&lt;sub&gt;A&lt;/sub&gt; Complexity</td>
<td>Levels 1 &amp; 2</td>
<td>( F(1, 27) = 25, p &lt; .001 )</td>
</tr>
<tr>
<td></td>
<td>Levels 3 - 6</td>
<td>( F(3, 81) = 8.83, p &lt; .001 )</td>
</tr>
<tr>
<td>IV&lt;sub&gt;B&lt;/sub&gt; Complexity</td>
<td>Levels 1 &amp; 2</td>
<td>( F(1, 27) = 25.6, p &lt; .001 )</td>
</tr>
<tr>
<td></td>
<td>Levels 3 - 6</td>
<td>( F(3, 81) = 7.13, p &lt; .001 )</td>
</tr>
</tbody>
</table>

Table 6.9. Mean performance levels by the complexity level of each IV. Standard errors are given in parentheses.

<table>
<thead>
<tr>
<th>Accuracy Across Statements</th>
<th>M (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IV&lt;sub&gt;A&lt;/sub&gt; Complexity</th>
<th>Level 1</th>
<th>.84</th>
<th>(.03)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Level 2</td>
<td>.72</td>
<td>(.03)</td>
</tr>
<tr>
<td>Level 3</td>
<td>.72</td>
<td>(.02)</td>
<td></td>
</tr>
<tr>
<td>Level 4</td>
<td>.66</td>
<td>(.02)</td>
<td></td>
</tr>
<tr>
<td>Level 5</td>
<td>.67</td>
<td>(.03)</td>
<td></td>
</tr>
<tr>
<td>Level 6</td>
<td>.61</td>
<td>(.02)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IV&lt;sub&gt;B&lt;/sub&gt; Complexity</th>
<th>Level 1</th>
<th>.84</th>
<th>(.02)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Level 2</td>
<td>.72</td>
<td>(.03)</td>
</tr>
<tr>
<td>Level 3</td>
<td>.71</td>
<td>(.02)</td>
<td></td>
</tr>
<tr>
<td>Level 4</td>
<td>.67</td>
<td>(.03)</td>
<td></td>
</tr>
<tr>
<td>Level 5</td>
<td>.67</td>
<td>(.02)</td>
<td></td>
</tr>
<tr>
<td>Level 6</td>
<td>.62</td>
<td>(.02)</td>
<td></td>
</tr>
</tbody>
</table>

Note: In essence, the pattern of means is the same for both IVs, although the patterns of significance vary. For both IVs, Level 2 took significantly longer than Level 1. For graphs with interactions, pairwise contrasts indicated that Level 6 was significantly harder than all other levels, and level 3 was significantly easier (marginally for IV<sub>A</sub>) than the other levels. The values are collapsed across the other dimension of the graph space. These accuracy values are estimated at the mean-centered Math Composite level.
In addition, this trend of increasing difficulty as complexity of the graph increases shows up within the levels of the different variables (See Table 6.10). The effects of the other variable were strongest when the variable of interest was at Level 6 and weakest when it was at Level 3. Increasing the complexity of IV_A significantly affected performance for level 6 of IV_B, and marginally affected performance at Levels 4 and 5 of IV_B. Increasing the complexity of IV_B significantly affected performance for levels 4 and 6 of IV_A. The effects of the two IVs were consistent when graphs did not have interactions. Simple effect analyses from the logistic measures analyses on the accuracy data indicated that the increasing complexity of IV_A negatively affected both levels of IV_B and the increasing complexity of IV_B negatively affected both levels of IV_A for graphs without interactions ($p$s < .01).

Table 6.10. Accuracy levels based on graph pattern for each graph from the simple effects consistency model. The overall effects seen in comparing the complexity levels of each IV are also represented by the individual graph patterns. Decreases in accuracy as complexity increases occur in the levels for each IV.

<table>
<thead>
<tr>
<th>IV_A Complexity</th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
<th>Level 4</th>
<th>Level 5</th>
<th>Level 6</th>
<th>Weighted Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1</td>
<td>.91</td>
<td>.77</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.8</td>
</tr>
<tr>
<td>Level 2</td>
<td>.77</td>
<td>.67</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.7</td>
</tr>
<tr>
<td>Level 3</td>
<td>.74</td>
<td>.71</td>
<td>.71</td>
<td>.70</td>
<td></td>
<td></td>
<td>0.71</td>
</tr>
<tr>
<td>Level 4</td>
<td>.71</td>
<td>.69</td>
<td>.68</td>
<td>.60</td>
<td></td>
<td></td>
<td>0.67</td>
</tr>
<tr>
<td>Level 5</td>
<td>.73</td>
<td>.67</td>
<td>.67</td>
<td>.61</td>
<td></td>
<td></td>
<td>0.67</td>
</tr>
<tr>
<td>Level 6</td>
<td>.70</td>
<td>.59</td>
<td>.63</td>
<td>.54</td>
<td></td>
<td></td>
<td>0.62</td>
</tr>
</tbody>
</table>

| Weighted Average | 0.84 | 0.72 | 0.72 | 0.66 | 0.67 | 0.61 |
6.11 Results: Influence of Math & Statistics Background

For reaction time, several three-way interactions were present when graphs were at Levels 3-6: Stmt x IV\textsubscript{A} Complexity x Stats, $F(27, 702) = 1.66, p = .02, \eta^2_p = .06$; and IV\textsubscript{B} Complexity x IV\textsubscript{A} Complexity x Stats, $F(9, 234) = 2.14, p = .03, \eta^2_p = .08$. In addition, Stmt x IV\textsubscript{B} Complexity x Math was significant, $F(27, 702) = 1.57, p = .04, \eta^2_p = .06$. These yielded no discernible pattern. The main effects of math and statistics were not significant ($p > .8$).

However, math and statistics background did affect accuracy of responding. On graphs without interactions, participants with statistics experience ($M = .84, SE = .03$) responded more accurately than those without ($M = .72, SE = .03$), $F(1, 24) = 8.02, p = .009$. Similarly, on graphs with interactions (levels 3 to 6), there was a main effect of statistics, $F(1, 24) = 9.17, p = .006$. As with the graphs without interactions, participants with statistics experience ($M = .73, SE = .03$) were more accurate than those without statistics experience ($M = .61, SE = .03$). (Note, however, that the accuracy levels still reflect the overall trend of increased accuracy with graphs without interactions compared to graphs with interactions.)

Statistics experience interacted with the complexity level of IV\textsubscript{A} when working with graphs at Levels 1 and 2, $F(1, 27) = 7.31, p = .012$. Simple effect follow-ups indicated that the beneficial effect of statistics occurred only for the first level of IV\textsubscript{A}.

For graphs at Levels 3 – 6, IV\textsubscript{B} (not IV\textsubscript{A}) influenced the effect of statistics experience, Statistics x IV\textsubscript{B} Complexity: $F(3, 81) = 4.22, p = .008, \eta^2_p = .12$. Participants having taken statistics were affected by the complexity of IV\textsubscript{B}, while participants without statistics
were not (see Figure 6.15). Simple effect analyses confirmed that there was no effect of IVB Complexity for participants who had not taken statistics, $p > .5$. The simple effect of IVB Complexity was still present for those who had taken statistics, $F(3, 81) = 10.62, p < .001$. In addition, the main effect of statistics (with higher performance if you had taken the course) showed up only in levels 3 – 5 of IVB, and not in Level 6.

Statistics interacted with the statement type for graphs at levels 3 – 6 (Levels 1 – 2: $p = .09$, Levels 3 – 6, $F(9, 243) = 2.83, p = .004$).

![Figure 6.15](image.png)

**Figure 6.15.** Statistics experience determined effect of IVB Complexity on performance. Participants who had previous statistics experience performed significantly better (across statements) than those who had not taken statistics. In addition, only those who had taken statistics were affected by the complexity level of IVB for graphs at Levels 3-6.
6.12 Results: Interactions Among Statements and Graphs

6.12.1 Interactions between Statement and Graph on RT

In the repeated measures analyses of RT for graphs at complexity levels 1 and 2, there was a Statement x Row interaction, $F(6.7, 175) = 2.33, p = .029, \eta^2_p = .08$. The interaction between statement and row is evident in the higher RT for some of the statements on row 2 compared to row 1. In particular, the two MC interaction statements and the MC statement about the main effect of IV_B took over 3 seconds longer when paired with graphs from IV_B level 2 than with graphs from IV_B level 1 (see Figure 6.16). The other statements showing increases also focused on the role of IV_B in the graph. Looking at the means more closely, the increase in RT is even more pronounced when both IVs are at the 2nd complexity level relative to the 1st; however, there was not a significant Statement x Row x Column interaction ($p > .4$). For graphs at Levels 3 – 6, Statement had a marginal interaction with IVA Complexity, $p = .058$. The effect of increasing complexity for IV_A led to increased RT for the directional main effect statement about IV_A.
6.12.2 Interactions between Statement and Graph on Accuracy

Not only were there effects of graph pattern on complexity and of statement on complexity, but the effects of graph pattern and statement interacted with one another. Simple effect analyses showed that the type of statement impacted performance for each level of IV_A and of IV_B, ps < .001. However, only some of the statements were affected the complexity of the graph pattern. Table 6.11 presents the inferential statistics, and Table 6.12 reports which statements were affected by the changing complexity level for each IV.
Table 6.11. Inferential statistics for IV Complexity x Statement interactions.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Graph Pattern in Analysis</th>
<th>Inferential Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>IV_A Complexity x Statement</td>
<td>Levels 1 &amp; 2</td>
<td>$F(9, 252) = 3.41, p &lt; .001$</td>
</tr>
<tr>
<td></td>
<td>Levels 3 - 6</td>
<td>$F(27, 756) = 4.08, p &lt; .001$</td>
</tr>
<tr>
<td>IV_B Complexity x Statement</td>
<td>Levels 1 &amp; 2</td>
<td>$p = .4$</td>
</tr>
<tr>
<td></td>
<td>Levels 3 - 6</td>
<td>$F(27, 756) = 2.31, p &lt; .001$</td>
</tr>
</tbody>
</table>

From this data, one can see that performance on directional interaction statements was affected by both IV_A and IV_B for graphs with and without interactions. The non-directional statements were not similarly affected, especially for IV_A. Performance on statements about simple effects was not affected by graph complexity. Changing the level of IV_B complexity (for levels 1 – 2 and levels 3 – 6) affected performance on both the directional and non-directional statements about the main effect of IV_B. Similarly, changing the level of IV_A complexity tended to affect performance on both the directional and non-directional statements about the main effect of IV_B. That is, the influence of complexity level was tied to the statements about the IV that was changing its complexity.
Table 6.12. Effects of complexity level by statement (IV, directionality, and statement type).

<table>
<thead>
<tr>
<th>IV</th>
<th>Directionality</th>
<th>Type of Statement</th>
<th>Changing IV&lt;sub&gt;A&lt;/sub&gt; Complexity</th>
<th>Changing IV&lt;sub&gt;B&lt;/sub&gt; Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Levels 1 - 2</td>
<td>Levels 3 - 6</td>
</tr>
<tr>
<td>A</td>
<td>directional (MC)</td>
<td>simple effect</td>
<td>~</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>directional (MC)</td>
<td>simple effect</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>directional (MC)</td>
<td>main effect</td>
<td>*</td>
<td>**</td>
</tr>
<tr>
<td>B</td>
<td>directional (MC)</td>
<td>main effect</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>non-directional (T/F)</td>
<td>main effect</td>
<td>***</td>
<td>*</td>
</tr>
<tr>
<td>B</td>
<td>non-directional (T/F)</td>
<td>main effect</td>
<td>***</td>
<td>**</td>
</tr>
<tr>
<td>A</td>
<td>directional (MC)</td>
<td>interaction</td>
<td>**</td>
<td>***</td>
</tr>
<tr>
<td>B</td>
<td>directional (MC)</td>
<td>interaction</td>
<td>***</td>
<td>***</td>
</tr>
<tr>
<td>A</td>
<td>non-directional (T/F)</td>
<td>interaction</td>
<td>?</td>
<td>*</td>
</tr>
<tr>
<td>B</td>
<td>non-directional (T/F)</td>
<td>interaction</td>
<td>?</td>
<td>**</td>
</tr>
</tbody>
</table>

~ p < .06   * p < .05   ** p < .01   *** p < .001

6.13 Combining and Summarizing Accuracy and RT Measures

Accuracy and RT were both used to evaluate the role of graph pattern and statement on participants’ graphical literacy. While there are similarities across the two dependent measures, they make slightly different distinctions in which levels of complexity result in significantly different performance.
Of critical importance here, is that the fact that not only do the two dependent measures generally line up with one another, but the dependent measures also parallel each other for both of the IVs. That is, not only is the ranking pattern the same for the complexity levels of IV\textsubscript{A} and IV\textsubscript{B}, but the actual values on the dependent measures are similar.

6.13.1 Effects of Statement

Overall, simple effects were easier than main effects, which were easier than statements about interactions. This general ranking of statement difficulty applied to graphs without interactions (complexity Levels 1 - 2) as well as graphs with interactions (complexity levels 3 - 6). Multiple choice (directional) statements were responded to more quickly than the T/F statements (non-directional). Accuracy was higher on directional versions for the main effects, but for the interaction statements, accuracy was higher on the non-directional versions. The particular IV referenced in the statement did not generally influence performance for graphs at either set of complexity. When there were differences in accuracy on the nondirectional (T/F) main effect statements, the statement about IV\textsubscript{B} was more difficult than the one referencing the main effect of IV\textsubscript{A}. In addition, the nondirectional (T/F) interaction statement was more difficult when it was focused on IV\textsubscript{B} rather than on IV\textsubscript{A}. 
6.13.2 Effects of Graph Pattern

$IV_A$ complexity had more of an effect on performance than $IV_B$ complexity. Levels 3 and 4 were less difficult than 5, but level 6 did not differ from any of the other levels. The most differentiation came for Levels 3 (it was easier) and Level 6 (it was harder). The general trends for accuracy were similar to those for RT overall.

Comparing the model to the traditional classification model, it is clear that the simple effects consistency model notices nuances that were not being distinguished by the traditional graph classification scheme. Figure 6.17 highlights the graph patterns that would be grouped together under the traditional classification. All of the graph patterns can be categorized according to the traditional classification scheme. In doing so though, one notes that performance is not consistent within each of those groupings. For instance, at Level 6 for either $IV_A$ or $IV_B$, there is considerable variation in performance levels as the level of the other variable changes. Thus, it appears that the influence of graphs with one main effect and one interaction on performance depends on the nature of the data pattern in the other variable.
Figure 6.17. Mean accuracy levels from current study compared to traditional classification. The prototypical patterns are indicated with dark green boxes, but as they are based on the number of effects, the other graphs delineated from the simple effects consistency model can be combined into the relevant pattern of the traditional model. However, as can be seen, particularly for Level 6 for both IVs, there is considerable variation within the pattern groupings delineated in the traditional classification scheme.

It is of interest to note, however, that the parallels between the complexity of IV_A and IV_B can be taken to indicate that the location of the IV in the graph doesn’t affect performance overall. This is unexpected because the two IVs produce geometrically distinct patterns. Thus, perhaps the conceptual nature of the relationship is what is influencing performance.

6.14 Are these patterns present in the individual participants?

Given the influence of statistics background on whether or not complexity influenced performance, as well as the variation across statements, it is particularly
important to consider whether or not individual subjects pattern in the same way as the group means. In order to test this, using the cell-based group of participants, the marginal means for each level of $I_{VA}$ and $I_{VB}$ Complexity were computed along with the marginal means for each statement. The means at the group level were then ranked. The same marginal means were computed for each participant, and the rank within each set was determined. Kendall’s coefficient of concordance ($W$) was used to determine the amount of variation in the ranks across subjects. In all cases, agreement was based on $p < .05$ for Kendall’s $W$.

For the directional (MC) statements about interactions and the non-directional (TF) statements about the main effect of $I_{VA}$ and of $I_{VB}$, the ranking of both $I_{VA}$ and $I_{VB}$ complexity levels of the individual subjects matched the overall means. Additional statements, such as the directional main effect for $I_{VB}$ also had similar rankings of the complexity levels of $I_{VA}$ and $I_{VB}$ within the individuals. On other statements, participants at the average and higher math levels were more likely to have rankings that agreed with the overall pattern.

Thus, in general, the degree to which individual subjects agree with the overall ranking appears to depend on the statement being evaluated. The effects of complexity level on interaction statements and main effect statements were more likely to be present in the individual data than for the other statements. The other statements are the simple effects – on which the high levels of performance could easily conceal the influence of IV complexity.
6.15 Summary and Discussion

In general, participants found various graph interpretations more difficult to evaluate when the simple effects depicted in the graph pattern were less consistent, that is, when the graphs were at a higher level of complexity as measured by the simple effects consistency model. While there is this general support for the model, however, it is also clear that the systematic effects of complexity were not present for all of the different statement types or between each complexity level. Thus, while there are also overall effects of statement type, the interaction between statement and graph pattern is critical to consider. Recognizing conceptual differences that underlie the varying patterns of performance will allow researchers to develop further understanding of the processes and strategies individuals employ when interpreting graphs.

In particular, if certain statements are hard to respond to when paired with certain graphs, is it the statement or the graph pattern that is responsible? Narrowing in on an answer to this question will allow researchers and educators to identify whether conceptual or procedural components are problematic. For instance, if one has limited understanding of interactions, if shown a graph that does not fit within your knowledge of interactions, you would not be able to interpret the statement correctly. In that case, the problem lies with the conceptual content. On the other hand, one could have a full understanding of interactions, but have difficulty using the representational features of the graph to assess that information. In this case, procedural aspects or other parts of graphical literacy are the source of difficulty.
In this task, the repeated measures of graph pattern and statement (crossed with different pairs) allowed the assessment of whether it is the particular graph pattern that poses difficulty across all statements, a particular statement that poses difficulty across all graph patterns, or, perhaps the more obvious answer, an interaction between the graph patterns and possible statements. This last situation was what occurred. In fact, there are suggestions that it occurred in a systematic way.

Interactions, especially directional statements, were influenced by the complexity of either IV. On the other hand, accuracies on simple effect statements were not affected by the complexity level of either IV for graphs with or without interactions. High levels of performance might have prevented this, but in addition, it is possible that comparing adjacent magnitudes, especially for IV_A comparisons, was an easy enough task such that graph complexity did not impact performance. For the main effects, the IV mattered. When IV_A was changed complexity levels, statements about the main effect of IV_A were affected. When IV_B changed complexity levels, statements about the main effect of IV_B were affected. This is an important finding as it can get lost in the overall picture of the difficulties of different statements and graph patterns. People appear to be differentially affected in terms of reasoning about a variable when that variable is the one that is changing.

One of the surprising findings is the similarity in the effects of the levels on both RT and accuracy as the complexity of IV_B and IV_A changes. This is a particularly important finding as it can be used to address the question of whether the effects are arising from conceptual or perceptual processes. The transitions between Levels 4 to 5
or Levels 5 to 6 are geometrically different for each IV, but they have the same underlying conceptual structure for both IV_A and IV_B. Thus, the similarity in the changing performance as the graphs went from level to level strongly suggests that conceptual characteristics of the data pattern are playing a role.

Presently, there is not enough evidence to fully identify the role of the multiple conceptual, perceptual, and procedural components that simultaneously interact, but there is enough evidence to suggest that working with these types of statements and systematic variation in graph patterns is a way to reach an answer to this question. Microgenetic studies that look at participants’ on-line reasoning as they work through these types of graph-statement pairings would be beneficial here. Theoretically, the approach suggests that the changes in the data pattern (either at the conceptual level or the perceptual level of the graph representation) interact with different types of statements – the interpretive task.

Looking at the different statement types, determining simple effects (in which two points on the graph are read and compared) were the easiest questions, while main effects and interactions were more difficult. These latter effect types require processes that are more complex in order to reach a solution. This is similar to the pattern of responding seen with the statements in the previous two studies. When there were differences among statements with the same content, they tended to be in favor of the statement involving IV_A. That is, the statements focusing on IV_B, were more difficult if there was a difference based on content type.
For main effect statements and for interaction statements, T/F statements took less time than MC statements for the same type of statement. These differences in reaction time could arise from a) the greater length or syntactic complexity of the MC statements and their response choices or b) the type of relationship specified by the statement. While the former possibility likely plays a role, it is the latter possibility that is of more interest. The T/F statements all presented non-directional versions of the statement. That is, the T/F statements only asked whether a particular effect was present or not. However, in the MC statements, the participant was asked to specify the direction of the effect. It is possible that it takes longer to identify the direction of the effect as compared to a detection model. However, this possibility cannot be distinguished from statement length or the number of responses in the current study.

Overall accuracy levels remain a concern. Future studies could benefit from addressing the role of feedback on performance. Investigating the effects of the different graph patterns and statements moves researchers closer to understanding how the strategies participants are using affect their performance when reading graphs. Understanding the reasons why certain statements and graph patterns interactively influence performance will help researchers identify the underlying processes. Developing a classification of the different graph patterns, such as the simple effects consistency model, adds greater structure to the research.

49 Of interest to this issue would be the reaction times to the same statement for graphs for which different response choices applied. However, the confound between graph pattern and correct answer, along with the effects of graph pattern in general, prevents a straightforward comparison of this issue.
CHAPTER 7

GENERAL DISCUSSION

Graphical literacy involves both the construction and interpretation of graphs. In the research studies presented here, both aspects of graphical literacy were addressed. In the graph construction task (Chapter 4), participants constructed a graph based on descriptions of the relationships among variables. The second study (Chapter 5) investigated the interpretation of statements about graphs and considered the effects of a statistics course on performance. The final study (Chapter 6) systematically manipulated the data pattern and paired it with various statements to provide a more nuanced consideration of how data patterns in graphs interact with the type of interpretation task.

Tying the studies together is the common theme of graphs with two independent variables (IVs). When two IVs are present, there is the potential for interaction. In Chapter 2, a novel classification scheme of graphs with two IVs was developed based on the consistency of the simple effects. This classification scheme was used to select the graph patterns used as stimuli. A second theme connecting the studies is the use of interpretation statements based on the effects of the IVs and their relationships with each other and the DV. Thus, all three studies involved statements about the main effects, simple effects, and interactions of variables in a graph.

The studies offer critical, new evidence about the concept of interactions and graphs with two IVs. The statement-specific effects support the general assertion that interactions are hard. Beyond that, however, the nature of the interaction has a
systematic effect on reasoning about a given statement. As the interaction patterns became more complex, participants' accuracy decreased and their RT increased. This might seem a self-evident statement. However, the key feature is that the complexity was increasing in a systematic way that has not been described and tested previously. This description represents an important advance for the research literature on graphing.

7.1 Assessing the Simple Effects Consistency Model: Effects of Graph Pattern on Graphical Literacy

Building up to the more systematic examination of graph pattern and statement interactions in Chapter 6, the studies presented support the validity of a more nuanced graph classification scheme. Performance variations based on the graph pattern allow for finer differentiations than are available in the traditional scheme of six or eight basic patterns (see Figure 2.6).

The simple effects consistency model (see Figure 2.11) was based on the conceptual relationships between the simple effects of an IV and the main effect of that IV. In systematically manipulating how consistent the sign and direction of the simple effects were, there are also systematic variations in the perceptual characteristics of the graphs at each level of complexity. Thus, one of the key questions for future research is the relative contributions of the conceptual or perceptual features to the changes in performance. Both are likely at play. Based on the Chapter 6 findings, there is strong evidence for a conceptual role as performance based on \( IV_a \) complexity levels mirrored that of \( IV_b \). This occurred despite the differing geometric or perceptual characteristics.
of the graphs. Thus, conceptual structure underlying the data pattern could be systematically influencing performance. (Two additional ways of addressing the relative contributions of conceptual and perceptual components will be considered below: investigating geometric transformations of data patterns (section 7.1.1) and assessing the role of these data patterns in non-graphical contexts.

Stepping back to the differences in performance that were found, the current research provides clear evidence to support the statement that graphs with interactions are more difficult to interpret. Putting the findings together, the general conclusion is that graphs that had data patterns with higher complexity levels had lower accuracies and longer RT than data patterns with lower complexity levels. However, this is qualified by several features – most notably, an interaction between the type of interpretation statement and the graph pattern. In addition, previous math and/or statistics experience affected not only performance levels overall, but also the impact of the data pattern’s complexity on performance. Thirdly, within the statements for which complexity of the data pattern did have an effect, not all levels were distinct from one another.

Somewhat surprisingly, given the findings that interpretation statements about IV_A and IV_B have differential effects on performance (Carpenter & Shah, 1998), complexity levels did not show this effect. That is, performance at a given complexity level of IV_A (collapsing across the complexity levels of IV_B) mirrored the performance at that same complexity level of IV_B (collapsing across the complexity levels for IV_A). This parallel occurred for both accuracy and reaction time.
In addition, results from Experiments 1 and 2, while not systematically varying the graph patterns, at the least does not contradict the theorized role of graph complexity on performance. That is, in the Construction task, accuracy was easier on graphs that had relatively less complexity. The data from the Statistics course study is not as clear cut, but it was clear that the graph pattern influenced performance.

7.1.1 A Comment on Categorizing Graph Patterns

When the different categorization schemes were introduced in Chapter 2, one of the general assumptions was that a single graph could represent a family of graphs that were related by geometric transformations. While the simple effects consistency categorization accounts for more variations in the relationship between the IVs, it still makes this assumption. In fact, it makes the assumption explicit by clearly defining the constraints underlying the generation of graph stimuli and the related data patterns that are not included. These related data patterns are part of the family of graphs (see Figure 2.11) that exist when one either geometrically or conceptually transforms the data. Having identified the assumptions and constraints used in generating stimuli, each of these aspects can be investigated. Part by part, a determination can be made about what types of transformations affect performance. This is working from the bottom-up to address what types of variations on graph pattern are empirically similar. It can be combined with the literature on graph perception for a top-down focus.

Parts of these assumptions have been tested in the research. For example, Hurts (2009) considered both graph patterns that arise when the placement of the IVa levels
are switched (i.e., changing the direction of the slopes), but he did not indicate that performance on the patterns systematically differed. Overall, this is still an open empirical question. It is possible that graph readers will treat the results of some, but not all, geometric transformations as analogous to the original graph. For instance, reversing the levels of IV_A may not influence performance. However, it is already clear that switching the locations of IV_A and IV_B does affect performance (Carpenter & Shah, 1998). Thinking of these graphs as being related both conceptually and geometrically helps organize the different transformations that are possible. Switching the locations of the two IV and using that as the underlying graph to generate stimuli will lead to similar sets of data patterns – but different sets of conceptual relationships among the variables.

Future research is needed to address this question of geometric transformations. If some of the geometric transformations behave in the same way as the original graphs, then there is a greater likelihood that the complexity of the data pattern is at a conceptual level, rather than a perceptual level. Looking at geometric transformations provides a way to start pulling this apart.

### 7.2 Effects of Interpretation Statements on Performance

Across the three tasks, general themes emerged about which statements were the most difficult. As predicted, simple effects were the easiest (at times, at ceiling), followed by main effects, and then followed by interactions. In addition, several of the results indicated that statements focused on IV_B tended to result in higher levels of
difficulty than statements focused on IV_A. In the construction task, simple effects were
more accurately represented than main effects; interactions could not be assessed.

Administering a graph-statement evaluation task at the beginning and end of a
statistics course (Chapter 5) allowed for an investigation of participants’ knowledge
before and after becoming more familiar with the content of the interpretation
statements. The participants significantly improved their performance on statements
about main effects as well as statements about interactions. However, there was still
substantial room for improvement in their overall accuracy levels. Across tasks, with the
exception of simple effects for some of the graph patterns, performance was not at
ceiling.

7.2.1 Types of Interpretation Tasks

The graph-statement evaluation tasks presented in Chapters 5 and 6 had a
constrained set of possible interpretations. The interpretation statements selected for
these experiments closely parallel the types of interpretations made in analyzing
research data. The emphasis was placed on the statistical ideas of main effects, simple
effects, and interactions. These conceptual categories encompassed a way to examine
the different manners in which two IVs could relate to one another.

However, it is also important to consider how focusing on the conceptual
content of the statement relates to other schemes that categorize types of
interpretation tasks. Using the statistical effects to categorize the interpretation
statements does not leave behind the types of skills or processes that are necessary to
obtain the content. Instead, the different approaches can act in concert. Friel et al.’s (2001) levels of interpretation are frequently referred to in considering what level of difficulty an interpretation task poses.

Simple effect statements are operating at least at Friel et al.’s (2001) second level of interpretation – table look-up is required as well as comparison. Initially, the interpretation task requires the graph reader to focus selectively on two of the magnitudes: either adjacent bars at one level of IV_A or bars sharing the same legend identification (i.e., at the same level of IV_B). It subsequently requires the graph reader to apply the information gained from reading the legend. Only then does the interpretation task come back to the table look-up values for each quantity. Finally, one must compare the two values obtained from the graph. Comparing main effects adds the additional requirement of averaging the bars across the levels of the other variable. This adds multiple steps into the process, again placing these types of statements at the second or possibly third level of interpretation. Finally, statements about interactions require both more operations with distinct data points on the graph as well as a higher working memory load in tracking the information about the comparisons.

Thus, all of these statements are operating in between the second and third levels of Friel et al.’s (2001) interpretation classification, but there are distinct differences in undergraduates’ performance. At this point, it is not clear whether these differences can be traced to the complexity of the processes underlying the interpretation of the statement or whether the conceptual content of the interpretation statement is affecting performance. Thus, again it becomes an issue of determining
what is driving the performance variations: the conceptual content of the statements, graphical literacy, or a combination of these two. Manipulations like that employed in the Construction task are necessary to separate out the relative effects of these components.

7.2.2 Other Types of Interpretation Tasks

The interpretations used in these two tasks were not open-ended. In contrast, when individuals encounter graphs outside of participating in an experiment, they are asked to generate their own interpretations along with evaluating others’ interpretations. In both of these cases, the interpretations do not belong to a constrained set, especially as the number of variables and their levels increase.

These non-constrained interpretations occur in a variety of modalities - in titles and captions as well as in text that refers back to the graph. One possible research direction would be to consider these more characteristic uses of graph interpretation as a measure of graphical literacy. Participants’ composition of titles or captions could demonstrate both the understanding of the graph content and more general graphical literacy skills relating to these elements of graphing conventions. The text in the graph captions or titles can reflect the different levels of interpretation that a graph reader pursues. However, most experimental and psychophysical studies have not asked participants to generate these texts or to work with texts along with graphs despite the typical activity of writing about graphs, such as writing summaries of data, in academic settings.
One of the difficulties with employing open-response tasks in this situation is that participants do have free choice of which strategies they use. One can determine what level a participant is operating at in their response, but participants can focus on different parts of the graph in their interpretation. They could, for instance, focus on individual data values or combine information across multiple categories, a higher level interpretation task. While this approach may indicate their preferred level of interaction with the given graphical representation, it is not necessarily an indication of the highest strategy available to a given participant.\textsuperscript{50} It is expected, however, that participants will make use of the highest strategy required to answer a question that they have available to them. Thus, a caveat is required if a question could be fully answered at multiple levels of interpretation; in that case, the use of a lower level strategy does not indicate that a participant can only gather that information from the graph; other strategies might still be available.

This concern is, however, applicable to the construction tasks. As the participants fulfilled the constraint satisfaction component of the task, they were employing higher cognitive reasoning skills for at least part of the task. There is no reason to believe that they were operating at a lower level than they were capable of when constructing the graph representation.

\textsuperscript{50} Participants may choose to answer the question using the strategy that is easiest for them to deploy even if they are capable of higher level interpretation skills. (See Siegler’s remarks about strategy deployment and the overlapping waves model.)
7.3 Graph Construction

Asking participants to construct graphs based on a series of statements about the relationships among the variables is a novel task. The construction task was modeled from Mevarech and Kramarsky’s (1997) research. They used a single statement that referenced one continuous IV and one DV. A series of statements becomes necessary as soon as the number of variables (and levels) involved in the data increases.

In the construction task conducted here, a significant issue arose in that the graphical representation of the data (or at least, of the existence of relationships among variables) presented additional challenges beyond that posed by determining the particular data pattern. However, simply adding graph-relevant information and structure by labeling and spatially locating the variables on the axes vastly improved performance and partially carried over to the near transfer 3rd graph. This is seen both in the quantitative measures of which quantities were represented on the graph and the more qualitative evaluation of the overall representation. Particularly relevant here is that the manipulation targeted specific skills. In and of itself, the accuracy of representing relationships did not change because of the manipulation. The manipulation increased the existence of the representations, but participants were equally able to satisfy the constraints imposed by the series of statements.

This targeted impact of adding labels is also supported by the analysis of the atypical graphs some of the participants produced. A common error pattern was putting one IV on each axis and not including the DV on the graph, which resulted in a
representation that did not make use of the information in the statements. Other error patterns also suggested that a general difficulty was integrating the two IVs. Asking participants to construct graphs from a series of statements, as opposed to from a table of values, also points out the additional required component of negotiating how variables and their levels relate to one another. In the table of values, the structure of the variables and their levels is more evident. Another type of error produced was incomplete mixing of the levels of the variables. The Added Labels condition identified and separated the variables, thus helping with these troublesome parts of the representational task.

Given the immediate and drastic improvement present on the 2nd trial in the Added Labels condition, it is likely that this skill had been previously learned. Then the question becomes how and when the skill was lost. Individuals continue to see graphical representations of two or more IVs in their college courses as well as in the broader media. Another possibility is that the manipulation helped them generate the conventional representation because of its organization.

This type of task clearly identified differences in levels of graphical literacy among participants. Analysis of the error patterns also points out areas of difficulty that can be targeted for future intervention. One of these, the idea of crossing the levels of the two IVs, is critical for both conceptual understanding of variables and interactions as well as for graphical representation of two IVs. Other participants appeared to have an idea that a graph with multiple variables should have multiple bars, but they did not link this to the data or the relevant variables. The results from the Construction task
indicate that a simple manipulation suffices to help subjects cross the levels of the two IVs. However, whether the difficulty lies at an earlier conceptual stage and/or with graph representation is an issue for further investigation. As with the need to investigate different data patterns in non-graph domains, it also will be necessary to separate out the concept from the graphical literacy components here.

An alternate possibility comes from considering how the information was presented to participants. The introduction and description of the variables and background context for the data participants received were in a typical and conventional format. It is possible, though, that this was not supportive enough. Perhaps some of the subjects were not able to generate the underlying study design of two IVs crossed with one another. Clarifying this in the background text may be a way to address whether the difference lies at the representational level or at an earlier conceptual level. Using the construction task in a more experienced population (one in which subjects had more exposure to research studies and designs) would also be important. If this is beneficial, then a goal is to address how to help students become able to use the less supportive, but conventional, background descriptions.

### 7.4 Influence of Math Background

The math levels in these experiments were split at Calculus I – those at or above this level compared to those for whom their highest math course was Pre-Calculus. However, as the tasks themselves did not require this level of math, it seems more likely that the split is picking up on a group difference in overall mathematical ability,
preference for mathematical thought, or perhaps an indicator of how much exposure to mathematics had been sought by the participant. Even so though, despite the small range of variation in math levels, there were consistent effects of math level across tasks. What is most important, however, is the existence of interactions between math/statistics background and the other factors in the task. Participants at different math ability levels were differentially affected by the complexity of the data pattern in the Speeded Evaluation of Graph-Statement Matches and by the effect of trial in the Construction task. These differences between the math groups are important to keep in mind when continuing research on these topics.

It is important to note that when composite math variables were used, each individual component of the composite variable did not have the same effect on the DV. It is theoretically reasonable to expect an impact of the different math background variables (e.g., highest level of math taken, math SAT score, statistics experience, math/science major, numeracy score) to have an effect, but it isn’t clear how to directly measure which background construct directly influences graphing ability or how to measure it. Previous research on the impact of math on graphing literacy is mixed.

7.5 A Comment on Interpreting Interactions

Rosnow and Rosenthal write that interactions are the “universally most misinterpreted results in psychology” (1989b, p. 1282). However, their basis for making this statement is not a conceptual misunderstanding of how multiple IVs relate to one another and a DV. Instead, Rosnow and Rosenthal focused on the statistical
investigation of an interaction and the accompanying interpretation of group means.

Umesh, Peterson, McCann-Nelson, & Vaidyanathan (1996) attempt to indicate that the multiple uses of the term ‘interaction’ confuse the issue. This appears to be an ongoing debate that has been left at a stalemate.

Zuckerman, Hodgins, Zuckerman, & Rosenthal (1993) asked authors published in APA journals whether testing simple effects was the correct approach to interpreting interactions. Around 60% correctly indicated that testing simple effects was not the correct approach to interpreting interactions. However, Zuckerman et al. went on to note that among these participants who correctly responded ‘no’, their comments indicated that their answer of ‘no’ was not based on reasoning that reflected Rosnow and Rosenthal’s approach to understanding interactions. The remaining subjects were split between indicating that it was and indicating that it depended. Zuckerman et al. also note that some of the respondents were aware of Rosnow and Rosenthal’s argument, but indicated that they disagreed with it.

Zuckerman et al.’s solution to their question about interactions provides a clear statement of the argument:

“Interaction effects cannot be interpreted on the basis of comparisons between cell means (the so-called simple effects) because these means combine the effect of the interaction with the effects of rows and columns (the main effects). Stated differently, the main effects may contribute to the simple effects as much as or even more than the interaction does (Rosenthal & Rosnow, 1991; Rosnow & Rosenthal, 1989a). The meaning of the interaction is defined in terms of the interaction residuals (i.e., leftovers of the cell means after all lower order effects have been removed). Of course, comparisons between cell means may be
important in their own right. However, such comparisons do not reveal the pattern of the interaction.” (p. 53)

The latter part of Zuckerman et al.’s (1993) quote is important to the current studies. Comparing cell means (i.e., the values arising when one fully crosses the levels of each IV) may not speak directly to whether a main effect or an interaction contributes to the cell means. However, comparing cell means does enable one to make a statement about how the different levels of the variables are related to one another. This is the topic of interest in the current experiments – evaluating and interpreting how one level of an IV behaves when it is combined with other IVs. This type of interpretation – relating levels of variables to one another at a descriptive level – is separate from Rosnow and Rosenthal’s (1989) focus on where the effects originate. In fact, Umesh et al. (1996) note that graphing cell means helps communicate how the variables are related to one another. In addition, in his statistics textbook, Howell (2009) acknowledges Rosnow and Rosenthal’s perspective, but then goes on to focus on interpreting cell means as the way to understand interactions. In the current studies, the goal was to gain a descriptive understanding of how the values of the DV vary based on reading the information from a graph. This level of understanding is available from graphs of cell means.

7.6 Future Directions

Measuring the understanding of graphs with two IVs moves researchers towards the possibility of a quantifiable measure of graph complexity that derives from the conceptual relationships among the graphed variables. Having a metric that
differentiates graph patterns enables researchers to focus on how participants interact
with the different graph patterns. These findings about interactions represented in
graphs and in interpretation statements about graphs support the generalized
statements and related research in the inductive / inquiry learning field that the
different relationships among variables are of different difficulty levels. The
experiments presented here demonstrate that these difficulties persist when reading a
graph.

Further investigating the classification of graph patterns will enhance the
structure that guides research in this area. Other types of tasks can also target the
impact of the data pattern’s complexity. For instance, asking participants to recall a
graph’s contents (either verbally or representationally) requires them to reference the
data pattern that they obtained from the graph when they initially viewed it. Working
with interpretation statements relating to the presence of simple effects also imposes
an organization that enables a cataloguing of the possible results from a given graph.
These are the types of interpretation statements academic researchers make when
describing and interpreting data. In doing so, it is important to make sure that the
participants in these studies share a common language with the graph designers. For
instance, one can talk about the qualitative size of an effect (e.g., bigger or smaller).
However, the simple effects consistency model reminds us of another dimension: both
magnitude and direction are involved. So one key unanswered question in regards to
using statements about effects in interpretation tasks is whether or not participants
treat ‘effect’ as a concept that has two dimensions. Additional variations in the
statements (e.g., non-directional versus directional) also seemingly impact performance.

Several of the remaining questions address the issue of what stage or area (e.g.,
conceptual, procedural, graphical) these difficulties originate at. In all likelihood, the
difficulties exist at multiple stages. However, isolating the concept and investigating it
at each stage (as was done with the Added Labels manipulation in the Construction task)
will further researchers’ and educators’ understanding of how this content knowledge
develops. If the difference between the graphs is based on conceptual differences in
the relationships the graphs express, then educational experiences that deal with
interacting variables could beneficially influence graph interpretation.

Identifying the statement-specific and graph-specific effects will aid the
identification and enhancement of participants’ interpretation procedures. Recognizing
the cases for which strategies succeed and fail will allow a more thorough
understanding of the strategies. This research is critical as the emphasis continues to be
on enhancing literacy in the STEM fields of science, technology, engineering, and math
as well as for graphical literacy in general.
Appendix 1

Descriptive Statistics from Entire Sample

in Speeded Evaluation of Graph-Statement Matches

In order to be considered for inclusion in the descriptive statistics of the full sample reported below, it was necessary for a subject to have at least 60% of the 400 trials be valid. This requirement eliminated 6 subjects from the data set. Of the 47 remaining participants, an average of 84% (SD = 13%) of their trials were valid and within time limits, for an average of 335 trials per subject.

The demographic characteristics of the full sample are similar (see Section 6.6) to those reported for the subset for whom first trial data was available. The subjects included in the analyses reported in Chapter 6 were not significantly different on the background variables from entire sample (independent sample t-tests, ps > .8). In this full sample, 40% were in their first or second undergraduate year. For this group, 47% had a math/science major, 62% were in the high (>= 600) math SAT group, and 70% were at the higher level of math experience (at least Calculus I). Forty-seven percent of the participants had taken a statistics course.

Effects of Statement on RT

As mentioned in section 6.9.1, differences in statement length and syntactic complexity preclude in depth comparison of RTs across different statement types (see Figure A1.1). For both the main effects and the simple effects, the non-directional (T/F) statements were faster than their directional (MC) counterparts. The MC statements
about interactions were longer than the other statements. The directional (MC) statements about main effects were the next longest. The nondirectional (T/F) statements about main effects were responded to more quickly than the other statements.

![Effect of Statement on RT](image)

**Figure A1.1.** Reaction time across all subjects ($n = 47$) and all graphs based on statement. The non-directional (T/F) statements were faster than their directional (MC) counterparts.

*Note.* Red bars indicate MC statements, and blue bars indicate T/F statements.

**Effects of Graph Pattern on RT**

Generalizing from the means reported in Figure A1.2 and Table A1.1, the graphs at complexity Levels 1 and 2 were responded to the fastest. In addition, Level 4 for IV$_B$ (at Levels 3 and 4 for IV$_A$) also tended to be on the faster side. The graphs at IV$_A$’s 5$^{th}$ and 6$^{th}$ complexity levels took the longest. This is seen in both the means for the
individual graphs (when crossing $I_{VA}$ and $I_{VB}$) as well as for the marginal means for the complexity of each IV.

![Graph](image)

**Figure A1.2.** RT varied based on the level of graph complexity for the full sample of data.

**Table A1.1.** The reaction times (seconds) represented in **Figure A1.2** are shown below, along with the marginal means. The increasing RT accompanying Levels 5 and 6 can be seen in the marginal means as it was for the individual graph patterns (where it was more prevalent for $I_{VA}$).

<table>
<thead>
<tr>
<th>IV$_A$ Complexity</th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
<th>Level 4</th>
<th>Level 5</th>
<th>Level 6</th>
<th>Weighted Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1</td>
<td>8.4</td>
<td>10.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>9.4</td>
</tr>
<tr>
<td>Level 2</td>
<td>10.8</td>
<td>11.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>11.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IV$_B$ Complexity</th>
<th>Level 3</th>
<th>Level 4</th>
<th>Level 5</th>
<th>Level 6</th>
<th>Weighted Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 3</td>
<td>12.3</td>
<td>11.9</td>
<td>13.0</td>
<td>13.0</td>
<td>12.5</td>
</tr>
<tr>
<td>Level 4</td>
<td>11.8</td>
<td>11.3</td>
<td>13.0</td>
<td>12.9</td>
<td>12.3</td>
</tr>
<tr>
<td>Level 5</td>
<td>12.5</td>
<td>12.8</td>
<td>13.3</td>
<td>13.4</td>
<td>13.0</td>
</tr>
<tr>
<td>Level 6</td>
<td>12.6</td>
<td>13.2</td>
<td>13.0</td>
<td>12.2</td>
<td>12.8</td>
</tr>
</tbody>
</table>

**Weighted Average** | 9.6 | 11.1 | 12.3 | 12.3 | 13.1 | 12.9
Effects of Statement on Accuracy

Across all graphs, the proportion of correct responses to statements ranged from .55 to .83, as seen in Table A1.2. On the directional (MC) questions, determining simple effects was the easiest; the next easiest type were the main effect statements. Only the MC main effect statement about IV_A was harder than both of the simple effect statements. Both of the interaction statements (in which the participant was asked to compare the size of the simple effects) were the most difficult of the directional (MC) statements. In particular, though, the interaction statement that focused on the effect of IV_B was harder than the interaction statement that focused on the effect of IV_A.

On the non-directional (T/F) questions, the two interaction statements were different from one another, and the two main effect statements were also different from one another. In both cases, the statements that dealt with the effect of IV_A had better performance than the statements dealing with the effect of IV_B.

Thus, among the three more difficult statement types, those that focused on IV_A were easier than their counterpart that focused on IV_B.

Table A1.2. Statement type and accuracy levels (across participants and across graphs in complexity levels 3 to 6). Data are presented for all valid trials, n = 47.

<table>
<thead>
<tr>
<th>Effect Type Referenced in Statement</th>
<th>Syntactic Notation of Statement</th>
<th>Proportion Correct for all Valid Trials</th>
</tr>
</thead>
<tbody>
<tr>
<td>main effect of IV_A</td>
<td>IV_A1 ? IV_A2</td>
<td>.74</td>
</tr>
<tr>
<td>main effect of IV_B</td>
<td>IV_B1 ? IV_B2</td>
<td>.77</td>
</tr>
<tr>
<td>simple effect of IV_A</td>
<td>in IV_B2, IV_A1 ? IV_A2</td>
<td>.81</td>
</tr>
<tr>
<td>-----------------------</td>
<td>--------------------------</td>
<td>-----</td>
</tr>
<tr>
<td>simple effect of IV_B</td>
<td>in IV_A1, IV_B1 ? IV_B2</td>
<td>.83</td>
</tr>
<tr>
<td>Interaction: compare simple effects of IV_A</td>
<td>IV_A had a ? effect for IV_B1 than IV_B2</td>
<td>.61</td>
</tr>
<tr>
<td>Interaction: compare simple effects of IV_B</td>
<td>IV_B had a ? effect for IV_A1 than IV_A2</td>
<td>.55</td>
</tr>
<tr>
<td>main effect of IV_A (T/F)</td>
<td>DV depended on IV_A</td>
<td>.66</td>
</tr>
<tr>
<td>main effect of IV_B (T/F)</td>
<td>DV depended on IV_B</td>
<td>.56</td>
</tr>
<tr>
<td>Interaction (T/F)</td>
<td>effect of IV_A affected by level of IV_B</td>
<td>.69</td>
</tr>
<tr>
<td>Interaction (T/F)</td>
<td>effect of IV_B affected by level of IV_A</td>
<td>.61</td>
</tr>
</tbody>
</table>

A Standard error was .01 for all of the statements.

**Effects of Graph Pattern on Accuracy**

Collapsing across statements, graph accuracy ranged from .53 to .91 as seen in **Table A1.3**. Examining the weighted averages for the rows and columns, the accuracies are similar for increases in the complexity of IV_A and IV_B. In particular, both the pattern and the actual values mirror one another. In general, the pattern appears to be suggestive of the hypothesized pattern, in which more complex relationships among the levels of a variable lead to more performance difficulties. One exception to the generally linear trend between levels is the joint pairing of each IV’s Level 2 complexity, which seems to have taken a particularly drastic hit on performance. In addition, participants do not appear to be distinguishing Levels 4 and 5 from one another for either IV.

**Table A1.3.** Graph Accuracy (across participants and across statements).
<table>
<thead>
<tr>
<th>IV&lt;sub&gt;A&lt;/sub&gt; Complexity</th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
<th>Level 4</th>
<th>Level 5</th>
<th>Level 6</th>
<th>Weighted Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1</td>
<td>.91</td>
<td>.75</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.8</td>
</tr>
<tr>
<td>Level 2</td>
<td>.76</td>
<td>.65</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.7</td>
</tr>
<tr>
<td>Level 3</td>
<td>.77</td>
<td>.72</td>
<td>.70</td>
<td>.69</td>
<td></td>
<td></td>
<td>0.72</td>
</tr>
<tr>
<td>Level 4</td>
<td>.73</td>
<td>.69</td>
<td>.65</td>
<td>.62</td>
<td></td>
<td></td>
<td>0.67</td>
</tr>
<tr>
<td>Level 5</td>
<td>.71</td>
<td>.68</td>
<td>.67</td>
<td>.60</td>
<td></td>
<td></td>
<td>0.67</td>
</tr>
<tr>
<td>Level 6</td>
<td>.70</td>
<td>.58</td>
<td>.58</td>
<td>.54</td>
<td></td>
<td></td>
<td>0.60</td>
</tr>
<tr>
<td>Weighted Average</td>
<td>0.84</td>
<td>0.70</td>
<td>0.72</td>
<td>0.67</td>
<td>0.65</td>
<td>0.62</td>
<td></td>
</tr>
</tbody>
</table>

The graph pattern occurring when both IVs were at Level 6 was harder than most others (exceptions tended to be in Level 6 for at least one of the IVs). The graph pattern occurring when both IVs were at Level 1 was easier than all others. Turning back to the patterns that the graphs represented, neither Graph 1 nor Graph 20 showed main effects. Graph 20 had an interaction that cancelled out the main effects.
Appendix 2

Effect of Graph Pattern on Accuracy and RT for Each Statement

From looking over the patterns presented in the following graphs, it is clear that the type of statement impacted the effect of complexity on accuracy and reaction time – however, it is equally clear that it did not do so in a consistent way across statements. As presented in Section 6.9.5, some of the statements are affected by the complexity level of both IVs, other statements are affected by only one IV’s complexity level, other statements are influenced at only a subset of the levels, and yet other statements are not influenced at all.

Delving into the effect of statement further is an important avenue for understanding how statement and graph pattern interactively influence a person’s graph interpretation. In addition, determining a metric to measure the size of graph complexity’s effect on the dependent measure is a step for further research.

For each statement, the descriptive statistics depicting the effect of graph pattern on accuracy and then RT are presented based on the 27 subjects for whom first trial data was available for all graph-statement cells.
Figure A2.1. Graph complexity had systematic effects on accuracy as one increased the IV_A complexity level for levels 3 – 5 of IV_B. For RT, increasing complexity of IV_A was associated with longer response times for levels 3 – 5 of IV_B.
Figure A2.2. There is some suggestion that accuracy is lower at levels 5 and 6 for IV_A. In addition, RTs appear longer for levels 5 and 6 of IV_A.
Figure A2.3. Graph pattern does not appear to systematically affect performance for the simple effect statement at IV_{B2}. High levels of performance may preclude its showing.
Figure A2.4. Graph pattern does not appear to systematically affect performance for the simple effect statement at IV_{A1}. High levels of performance may preclude its showing.
Statement 5: $I_{VA}$ had a $? \times ?$ effect for $I_{VB1}$ than $I_{VB2}$.

Figure A2.5. Level 6 of $I_{VA}$ complexity as well as Level 6 of $I_{VB}$ complexity appear more difficult in terms of accuracy levels.
Statement 6: IV_B had a ? effect for IV_A1 than IV_A2.

RT on Statement 6: IV_B had a ? effect for IV_A1 than IV_A2.

**Figure A2.6.** For accuracy, there appears to be an effect of increasing complexity of IV_B. In addition, the complexity of IV_A appears to have a non-linear trend, in that level 6 is the hardest and level 4 also poses more difficulty than the remaining levels of IV_A. The patterns of IV_B do not replicate.
Statement 7: DV depended on $IV_A$

**Figure A2.7.** Accuracy decreased as the complexity level of $IV_A$ increased, for each level of $IV_B$. In addition, accuracy had a slight tendency to be lower at higher complexity levels of $IV_A$. Reaction time did not appear to be affected by graph complexity.
Figure A2.8. Graph complexity affected accuracy such that performance was higher on graphs at IV_A’s complexity level 3. For RT, the patterns did not replicate, although there is some suggestion of an effect of IV_A complexity at levels 5 and 6 of IV_B.
Statement 9: effect of $I_{V_A}$ affected by level of $I_{V_B}$

Graph Pattern (Complexity for $I_{V_A}$ nested inside $I_{V_B}$)

RT on Statement 9: effect of $I_{V_A}$ affected by level of $I_{V_B}$

Graph Pattern (Complexity for $I_{V_A}$ nested inside $I_{V_B}$)

**Figure A2.9.** Graph complexity had systematic effects on accuracy as one increased the $I_{V_A}$ complexity level for levels 3 – 5. For RT, the patterns did not replicate.
Figure A2.10. For accuracy (and RT), there appeared to have been a non-linear trend for IV_A Complexity at IV_B’s complexity levels of 5 and 6.
References


CV

Jennifer L. Cooper

EDUCATION

Rutgers University, Psychology, Ph.D. October, 2010

Rutgers University, Psychology, M.S., 2004


ADDITIONAL RESEARCH EXPERIENCE

Graphing (construction and interpretation) studies with adults 2007 – ongoing

Number and graphing studies with preschoolers 2002 – 2007

Preschool science program 2002 – 2005

Carleton College: History Department 2001

  Analyzed (using SPSS) Dr. Jamie Monson’s data on Tanzanian railroads transportation

Massachusetts Institute of Technology: Behavioral Neuroscience laboratory Summer 2000

  Tested older adults on source memory and executive function tasks.

  Analyzed and presented the results and integrated them with previous literature

TEACHING EXPERIENCE

Rutgers University
Cognitive Science Courses 2009 – 2010

Teaching Assistant / Recitation Leader, Advanced Topics in Cognitive Science

Teaching Assistant / Recitation Leader, Introduction to Cognitive Science

Infant and Child Development Lab 2007 – 2009

Head Teaching Assistant

Course Instructor


Course Instructor

Infant and Child Development (lecture course) Summer 2004, 2006

Course Instructor

Carleton College: Mathematics and Computer Science Department 1999 - 2002

Prefector. Teaching assistant / tutor for Intro & Intermediate Computer Science

Lab Assistant in Computer Science Lab

Grader for Computer Science and Calculus courses