

UNIVERSITY OF CALIFORNIA SAN DIEGO

**Tales of Graphical Discovery:
A Case Study at the Intersection of Graph Comprehension and Visualization Design**

A dissertation submitted in partial satisfaction of the
requirements for the degree

Doctor of Philosophy

in

Cognitive Science

by

Amy Rae Fox

Committee in charge:

Professor James D. Hollan, Chair
Professor William Bechtel
Professor Philip J. Guo
Professor Mary Hegarty
Professor David J. Kirsh
Professor Arvind Satyanarayanan
Professor Caren M. Walker

2022

Copyright

Amy Rae Fox, 2022

All rights reserved.

The Dissertation of Amy Rae Fox is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

DEDICATION

*For La Jolla Shores
on the historically unceded territory of the Kumeyaay people
on whose sandy breaks I long pondered
the practice of science, the cognitive sciences, design,
and triangular-shaped graphs.*

EPIGRAPH

You, who are blessed with shade as well as light, you, who are gifted with two eyes, endowed with a knowledge of perspective, and charmed with the enjoyment of various colours, you, who can actually *see* an angle, and contemplate the complete circumference of a Circle in the happy region of the Three Dimensions—how shall I make clear to you the extreme difficulty which we in Flatland experience in recognizing one another's configuration?

Edwin A. Abbott
Flatland. A Romance of Many Dimensions

The Way It Is

There's a thread you follow. It goes among
things that change. But it doesn't change.
People wonder about what you are pursuing.
You have to explain about the thread.
But it is hard for others to see.
While you hold it you can't get lost.
Tragedies happen; people get hurt
or die; and you suffer and get old.
Nothing you do can stop time's unfolding.
You don't ever let go of the thread.

William Stafford

TABLE OF CONTENTS

Dissertation Approval Page	iii
Dedication	iv
Epigraph	v
Table of Contents	vi
List of Figures	viii
List of Tables	xi
Preface	xii
Acknowledgements	xiii
Vita	xvii
Abstract of the Dissertation	xviii
Chapter 1 Introduction	1
1.1 Representation and the Distributed Cognitive System	3
1.2 Models and Theories of Graph Comprehension	15
1.3 Methods in Graph Comprehension Research	38
1.4 An Unconventional Graph	44
1.5 Research Goals	48
Chapter 2 Explorations of Explicit Scaffolding	52
2.1 Cognitive Aids for Graph Comprehension	52
2.2 Research Goals	57
2.3 Study 1: Observing Interaction With an Unconventional Graph	58
2.4 Study 2: Testing Explicit Scaffolds	68
2.5 General Discussion	81
Chapter 3 Explorations of Insight Problem Solving	83
3.1 Problem Solving and Insight	84
3.2 Research Goals	87
3.3 Study 3A: Testing the Mental Impasse	87
3.4 Study 3A: Online Replication	105
3.5 Study 3B: Implicit (vs) Explicit Scaffolding	107
3.6 Study 3C: The Role of Working Memory	117
3.7 General Discussion	129
Chapter 4 Explorations of The Graphical Framework	132

4.1	The Graphical Framework	133
4.2	Research Goals	137
4.3	Experiment 4A: Gridlines	138
4.4	Experiment 4B: Marks	146
4.5	Experiment 4C: Shape and Scale	151
4.6	Experiment 4D: Orientation	158
4.7	General Discussion	165
Chapter 5	Conclusion	169
5.1	Summary of Findings	169
5.2	Implications for Visualization Design	171
5.3	Implications for Research	172
5.4	A Personal Reflection, and Future Work	173
Appendix A	Study 1 & 2 Supplementary Material	175
A.1	Study 1 & 2 — Materials	175
A.2	Study 2 — Supplemental Results	182
Appendix B	Study 3A — 3C Supplementary Material	184
B.1	Study 3A — Lab Supplementary Material	185
B.2	Study 3A — Online Replication Supplementary Material	187
B.3	Study 3B — Supplementary Material	189
B.4	Study 3C — Supplementary Material	191
Appendix C	Study 4A — 4D Supplementary Material	193
C.1	Study 4A — Supplementary Material	194
C.2	Study 4B — Supplementary Material	195
C.3	Study 4C — Supplementary Material	196
C.4	Study 4D — Supplementary Material	197
Appendix D	Multiple Choice Multiple Answer Scoring Strategy	198
D.1	Response Encoding	199
D.2	Scoring Schemes	201
D.3	Deriving the Interpretation Measure	205
Bibliography	206	

LIST OF FIGURES

Figure 1.1.	A collection of visuospatial external representations	5
Figure 1.2.	The three components of a Peircean sign	8
Figure 1.3.	The Human-Information Interaction Epistemic Cycle	13
Figure 1.4.	Early theories, frameworks and models in Graph Comprehension	18
Figure 1.5.	Four contributions ranking perceptual accuracy of visual-spatial encodings.	20
Figure 1.6.	Schematic diagram of Simkin & Hastie's Elementary Mental Processes ..	24
Figure 1.7.	A process description of visual information processing.	25
Figure 1.8.	Three Information Processing accounts of Graph Comprehension.	27
Figure 1.9.	A Construction-Integration Model of Graph Comprehension.	31
Figure 1.10.	Allen's Interval Logic	46
Figure 1.11.	Two Graphs for Interval Relations	46
Figure 1.12.	Allen Relations in LM and TM Graphs	49
Figure 2.1.	Study 1 (Materials) Layout of the Graph Reading Task	59
Figure 2.2.	Study 1 (Results) Graph Reading Artifacts	61
Figure 2.3.	Study 1 (Results) An Orthogonality Bias	62
Figure 2.4.	Study 1 (Results) Graph Orienting Behaviour	64
Figure 2.5.	Study 1 (Results) Highlighting Axis Scaffolds	66
Figure 2.6.	Study 1 (Results) Text and Hybrid Scaffolds	67
Figure 2.7.	Study 2 (Materials) Scaffold Conditions	69
Figure 2.8.	Study 2 (Materials) Stimuli	73
Figure 2.9.	Study 2 (Results) Total Score	74
Figure 2.10.	Study 2 (Results) Non-Equivalence of Scenarios	78
Figure 2.11.	Study 2 (Results) Drawing Types	79

Figure 3.1.	Study 3A (Materials) Experimental Conditions	88
Figure 3.2.	Study 3A (Materials) Layout of Graph Interpretation Task.....	91
Figure 3.3.	Study 3A (Materials) Interpretation Measure	96
Figure 3.4.	Study 3A (Results) Distribution of Total Score	99
Figure 3.5.	Study 3A (Results) Accuracy	101
Figure 3.6.	Study 3A (Results) Mouse Cursor Behaviour.....	102
Figure 3.7.	Study 3A (Results) Interpretation	104
Figure 3.8.	Study 3A Online Replication (Results) Distribution of Total Score	107
Figure 3.9.	Study 3B (Materials) Experimental Conditions	109
Figure 3.10.	Study 3B (Results) Distribution of Total Score	113
Figure 3.11.	Study 3B (Results) Accuracy	114
Figure 3.12.	Study 3B (Results) Interpretation.....	116
Figure 3.13.	Study 3C (Example) Costly End Time Calculation	120
Figure 3.14.	Study 3C (Results) Distribution of Total Score	126
Figure 3.15.	Study 3C (Results) Accuracy	127
Figure 3.16.	Study 3C (Results) Interpretation.....	128
Figure 4.1.	Pinker's Information Flow in Graph Comprehension	134
Figure 4.2.	Study 4A (Materials) Grid Design Conditions	139
Figure 4.3.	Study 4A (Materials) Layout of Graph Interpretation Task.....	142
Figure 4.4.	Study 4A (Results) Distribution of Total Score	144
Figure 4.5.	Study 4A (Results) Accuracy	145
Figure 4.6.	Study 4B (Materials) Mark Design Conditions	147
Figure 4.7.	Study 4B (Results) Distribution of Total Score	150
Figure 4.8.	Study 4B (Results) Accuracy	150

Figure 4.9.	Study 4C (Materials) Shape & Scale Design Conditions	153
Figure 4.10.	Study 4C (Results) Distribution of Total Score	156
Figure 4.11.	Study 4C (Results) Accuracy	157
Figure 4.12.	Study 4D (Materials) Shape & Rotation Design Conditions	160
Figure 4.13.	Study 4D (Results) Distribution of Total Score	163
Figure 4.14.	Study 4D (Results) Accuracy	164
Figure A.1.	Study 2 (Results) Linear Model Graph Scores	182
Figure A.2.	Study 2 (Results) Paired Graph Scores	183
Figure D.1.	Sample MCMA Question	198
Figure D.2.	MC vs MCMA vs MTF Question Formats	200

LIST OF TABLES

Table B.1.	Study 3A (Lab) Question Accuracy	185
Table B.2.	Study 3A (Lab) Question Interpretation	186
Table B.3.	Study 3A (Replication) Question Accuracy	187
Table B.4.	Study 3A (Replication) Question Interpretation	188
Table B.5.	Study 3B (Explicit vs Implicit) Question Accuracy	189
Table B.6.	Study 3B (Explicit vs Implicit) Question Interpretation	190
Table B.7.	Study 3C (Working Memory) Question Accuracy	191
Table B.8.	Study 3C (Working Memory) Question Interpretation	192
Table C.1.	Study 4A Question Accuracy	194
Table C.2.	Study 4B Question Accuracy	195
Table C.3.	Study 4C Question Accuracy	196
Table C.4.	Study 4D Question Accuracy	197

PREFACE

This dissertation is structured as three series of empirical studies (Chapters 2, 3, and 4) contextualized with a general Introduction and Conclusion. The Introduction and Conclusion *chapters* contain information relevant to all studies, while the introduction and conclusion *sections* of each individual chapter contain information relevant specifically to that series of studies. Before reading Chapters 2, 3 or 4, readers are encouraged to first review Section 1.4 that introduces the strange but elegant and computationally efficient graph used as the underlying stimuli for this body of research.

Supplemental materials including stimuli, data, and analysis scripts (in R) for all studies described in this dissertation are available upon request, and a web companion documenting analyses can be found at: <https://amyraefox.com/dissertation>.

ACKNOWLEDGEMENTS

In *Laboratory Life* (1987), Bruno Latour and Steve Woolgar convincingly argue that science is best understood as a socially-situated practice, enacted through learned routines (mundane and otherwise) and the interaction between researchers within networks of institutional structures. *If it takes a village to raise a child, it takes a (research) community to raise a scholar.* This was true of my journey to becoming a Cognitive Scientist, which I was privileged to undertake at an institution with one of the world's first Departments of Cognitive Science: UC San Diego. This dissertation, and the personal and professional growth the work brought about, would not have been possible without the UCSD's vibrant community, and there are so many here to whom I owe enormous debts of gratitude.

I am grateful to my advisor, Dr. Jim Hollan, who gave me tremendous freedom to follow the questions that most ignited my intellectual passions. No matter how far afield I might have strayed into psychology or statistics or philosophy, Jim was always there to help me see (and insist that I find) *the bigger picture*. Jim instilled in me two values I trust will guide and enrich all my future scholarly endeavours: (1) that it is important to take the time to think big thoughts, even when there are no easy (or productively publishable) answers, and (2) if you care deeply enough about a phenomenon, you will take the time to observe it. I hope the future will allow us many more years of collaboratively falling down rabbit holes.

I am grateful for the support and commitment of time offered by my dissertation committee, Drs. David Kirsh, Philip Guo, Caren Walker, William Bechtel, Arvind Satyanarayan, and Mary Hegarty. They are role models for how to contribute work with theoretic depth and practical import, at intersection of multiple disciplines. I am honoured to have assembled such a brain trust to support and critique this work, and only wish I would have taken greater advantage of their expertise in the hurried and chaotic years of the pandemic.

I am grateful for the guidance of my second year project committee, Drs. David Kirsh, Rafael Nuñez and Seana Coulson. When I began my second year project (Chapter 2 of this dissertation) I did not expect it would turn *into* my dissertation. Much to the contrary, I thought that I was

an outlier in my interaction with the graph that forms this case study, and that my second year project would be a nice replication with a neat little ceiling effect. I could not have been more wrong, and am grateful for the ways that each member of this committee contributed to my re-conceptualization of the core research question. Thank you to Rafael, for encouraging me to explore the literature on mathematical behaviour (and independently inventing yet another novel representation of time intervals on the back of a post-it note, during my talk). To David: thank you for countless cups of tea, chats about Peircean semiotics, and the relationship between the cognitive sciences and design. To Seana: thank you for introducing me to Conceptual Integration Theory, and forever dooming me to seeing *every* act of graphing as a conceptual blend. I hope you will not be too disappointed that I continued to study these ‘silly little graphs’.

I owe a special debt of gratitude to Dr. Caren Walker for providing me a secondary home base in the Early Learning & Cognition Lab. She was a constant cheerleader for my ability to continue my work no matter what obstacles I was facing, and I am enormously grateful for her moral, intellectual and logistical support. That I happened to be taking her class on *Learning by Thinking* while puzzling over the surprising results of Study 2 was a marvellous coincidence that moved this work in a new and much more exciting direction. I promise I will not give up until these papers are published.

I offer thanks to all of the talented instructors I’ve had the privilege of working with in my (many, *many*) TA-ships at UCSD: Drs. Nadir Weibel, Jim Hollan, David Kirsh, Federico Rossano, Bob Glushko, Taylor Jackson Scott, Drew Walker, Christine Johnson, Mary Boyle, Eric Leonardis, and Judy Fan. I learned something unique about teaching from each one of you. I did not expect I would enjoy teaching so much, that it would generate so many research questions, or that it would become such a large part of my PhD journey. But I do, it does, and it did, and I hope to contribute to the scholarship of teaching and learning (especially with visuospatial representations) well into the future.

To Judy Fan and the members of the Cognitive Tools Lab, thank you for inviting me into your community to contribute to the important work of defining and measuring graph comprehension

(or graphical literacy, graphicacy, graph reading, graph interaction, or whatever we decide to call it).

Thank you to Drs. Marta Kutas and Andrea Chiba who, through facilitation of the second and third year project classes helped me hone the voice in my head that constantly whispers, “*Is this Cognitive Science?*” Whether it is or isn’t, with their influence I’m better able to co-construct knowledge across the subdisciplinary lines that make up our vibrant field(s), and to appreciate contributions at differing levels of analysis.

I offer thanks to the administrative staff of the Departments of Cognitive Science and Psychology at UCSD for support for tasks too numerous to name, but without with the research and learning doesn’t happen.

I offer thanks my PhD cohort, (Tricia, Shuai, Kevin, Michael, Reina and Eric) **and my adopted cohorts** (Joey, Eric, Arthur, Tom, Richard, Rachel & Carson; Celia, Tania, Parla, Pam, Sean & Julia) and friends in the CogSci and Psychology community (Tibbles, Rose, Adam, Melissa, Cameron, Oisin, Srishti, Sean, Hui Xin & Emilia, Will, Hannah, and Holly). Can you believe how far we’ve come? Most of the most valuable lessons I learned on this journey were with and from these talented scholars, learning about the boundaries of the cognitive sciences (what is cognitive science?) and how to effectively communicate across disciplinary spaces.

I offer thanks to my pandemic pod, Ailie, Chris, Tricia, Rob & Ariana: thank you for helping keep things in perspective, sandy if not sanitized , and (relatively) sane during those years in the upside down.

Thanks to the members of the UCSD Design Lab (especially Philip, Steven, Scott, Bill, Nadir, Lily, Michèle, Stephanie, Mayya, Sean, Sam, Tone, Srishti, Steven, Danilo, Janet, Akshitha, Dorothy, Teenah, Deborah, Lars, Colleen, Cat & Ailie) for lively multidisciplinary (mis)adventures, camaraderie, and dramatically reshaping my mental model of the word *design*.

To my dear friend and soon to be PhD, Lauren, thank you for reminding that as challenging as Bayesian multilevel modelling might seem, at least it is not medieval history.

To Carson & Rachel: *thank you. There are not enough words.* In some ways you are like some

other people, and in some other ways, you are not.

I owe the final and greatest thanks to my family: To Wayne and Cindy, for your endless love and support, hardwork lessons of the farm and making me do 4H public speaking which I'm convinced has largely enabled me to write and speak in a scholarly fashion. To Ashley and April, for all the things. To Marlyce, for being a constant inspiration and Clark for stubbornly believing in my ability to do anything I set my mind to (if I could only make up my mind on which to set it upon). To Justin for keeping me humble, and to Chelsie, Jon and Trevor for keeping Justin, April and Ashley sane. To Clara, Callen, Tuftin, Logan, Huxlin and Benjamin: now its your turn!

CHAPTER 1, in part, includes material accepted for publication in the following venues:

(1) Fox and Hollan (2023) *Visualization Psychology: Foundations for an Interdisciplinary Research Programme*. In *Visualization Psychology*, Springer. (2) Fox (2023) *Theories and Models of Graph Comprehension*. In *Visualization Psychology*, Springer. The dissertation author was the primary investigator and author of these chapters.

CHAPTER 2, in part, includes material as it appears in: Fox and Hollan, 2018. *Read It This Way: Scaffolding Comprehension for Unconventional Statistical Graphs*. In P. Chapman, G. Stapleton, A. Moktefi, S. Perez-Kriz, & F. Bellucci (Eds.), *Diagrammatic Representation and Inference* (pp. 441–457). Springer International Publishing. The dissertation author was the primary investigator and author of this paper.

CHAPTER 3, in part, includes portions of material as it appears in: Fox, Hollan, and Walker, 2019. *When Graph Comprehension Is An Insight Problem*. In Proceedings of the Annual Conference of the Cognitive Science Society. Additional material appearing in this chapter is being prepared for publication. The dissertation author was the primary investigator and author of these publications.

CHAPTER 4, in part, is currently being prepared for submission for publication of the material. The dissertation author was the primary investigator and author of this material.

VITA

2004	Bachelor of Science, Computer Science University of North Carolina at Chapel Hill
2015	Master Sciences de l'Education Université Pierre Mendès France
2015	Master of Arts, Interdisciplinary Studies — Cognitive Visualization California State University, Chico
2022	Doctor of Philosophy, Cognitive Science University of California San Diego

FIELDS OF STUDY

Major Field: Cognitive Science

Studies in Cognitive Psychology
Professors James Hollan and Caren Walker

Studies in Information Visualization
Professors James Hollan and Arvind Satyanarayan

Studies in Human Computer Interaction
Professors James Hollan and Philip Guo

Studies in Learning Science
Professors Erica de Vries, Wolfgang Schnotz, and Neil Schwartz

ABSTRACT OF THE DISSERTATION

Tales of Graphical Discovery: A Case Study at the Intersection of Graph Comprehension and Visualization Design

by

Amy Rae Fox

Doctor of Philosophy in Cognitive Science

University of California San Diego, 2022

Professor James D. Hollan, Chair

“A picture is worth 1000 words,” the adage goes, but only—I argue—if you know how to read it. The same is true of graphs, charts, and diagrams. As powerful as these visuospatial technologies may be in their communicative efficiency, they needn’t be immediately easy to understand. In fact, there are often trade-offs between a graph’s *discoverability* and efficiency. Even for informationally equivalent forms, the computational efficiency of a less conventional representation may outweigh its ease of use for the untrained reader. It is this fact that underlies much innovation in Information Visualization, and the development of sophisticated interfaces for highly skilled workers performing specialized tasks. Sometimes this work results

in novel, unconventional representations that are computationally suited to particular complex tasks, but that would present a substantial challenge to the novice reader. Meanwhile, most work in remediating errors in graph comprehension has focused on “second order” readings: characterizing the trends or relationships between data represented in a graph. The ability to make these readings allows us to use graphs as vehicles for learning concepts—especially in science. We tend to accept *a priori* that well-designed graphs readily afford first-order readings: operations for extracting data from a graph. Accordingly, we know more about learning *with* representations, than we do about the learning *of* representations.

In this dissertation, I use simple graphs with an unconventional coordinate system to explore *graphical discovery*: how readers extract information from a graph when they lack prior knowledge of its graphical formalism. I address what the systematic errors readers make can tell us about our graphical intuitions, and the interaction of perceptual and conceptual processing that underlies graph comprehension.

Chapter 1

Introduction

Have you ever been struck by the beauty of a figure? You come upon a complex and colourful graph teeming with information, surely important by way of the precious column inches it spans. But as you scan for patterns—willing the authors insight to leap off the page—you find there is something unattainable. Like the writing of a foreign language you see familiar symbols and structure, but the rules for how to assemble the pieces into a meaningful whole are just outside your grasp. How do you make sense of the information?

Researchers of learning and cognition have long been interested in how graphic displays are used to communicate, solve problems, and generate insight: as tools to communicate thought, and tools to facilitate thinking. Through *structure*—the arrangement of marks in space—external representations facilitate the communication of arbitrarily complex ideas (Kirsh, 2010; Scaife and Rogers, 1996). Yet their power goes beyond communication from one mind to another. We know that by externalizing our knowledge into the world, we can perform operations with it beyond what we might achieve ‘inside’ our minds alone. It is the structure of a representation that influences the kinds of operations that can be performed with it (Cheng, 2016; Larkin and Simon, 1987; Palmer, 1978). From this powerful insight we’ve designed new forms: diagrams, charts, and interactive graphics to help people learn about complex and abstract concepts as diverse as molecular models in chemistry (Stull, Barrett, and Hegarty, 2013), particle collisions in physics (Cheng, 1996), properties of electricity (Cheng, 2002), the

mathematics of conditional probability (Binder, Krauss, and Bruckmaier, 2015) and timescales in geology (Resnick, Newcombe, and Shipley, 2016). Beyond learning and instruction, there is evidence for the role of external representations in generating insights at the frontiers of scientific inference (Bechtel, Abrahamsen, and Sheredos, 2018; Gooding, 2010; Kaiser, 2005). External representations form part of a distributed cognitive system in which *thinking* is constituted both inside and outside the body, via interaction with the environment (Hutchins, 1995). The empirical evidence supports what we intuitively feel to be true: external representations help us understand things *differently*. They are truly an example of “things that make us smart,” (Norman, 1993.)

Yet as powerful as graphics may be in their communicative efficiency, they needn’t be immediately easy to understand. There are most often trade-offs between a representation’s discoverability (ease of *deriving* the rules of the representational system) and efficiency (ease of *applying* the rules to perform operations). Consider the case of written language. The most primitive forms of written language were pictorial, employing marks on surfaces that bore visual resemblance to their referents. To construct meaning from the logogram  one need only have some familiarity with water fowl. Contrast this with the knowledge required to construct the same meaning from the English word “duck”. The relationship between the orthographic letterforms and real world referent are arbitrary, must be established through cultural transmission and learned by the individual. The logogram is more discoverable: the intended meaning of the form is within the perceptual grasp of the reader. Yet although it is less discoverable, as a representational system the English language is exceptionally efficient, affording a vast communicative potential with only 26 symbols.

Like language, understanding a graph or diagram is an inevitably semiotic process, as we endeavour to construct meaning for a sign purposefully constructed by a fellow meaning-maker to refer to their interpretation of something(s) in the world: a game of semiotic telephone. But what if we don’t know the rules of the game? Even familiar representational systems like scatter plots and line graphs can prove challenging for students (Shah and Hoeffner, 2002) and experts (Roth, 2003) alike. In short, we know a great deal more about learning *with* representations than

we do about the learning of representations. As Larkin and Simon note in their seminal paper, “a representation is useful only if one has the productions that can use it,” 1987, pg. 71. **If we lack the ability to draw inferences from a representation, then we may find it largely useless.**

In this dissertation I build upon previous research on reading and graph comprehension to explore how readers make sense of a particular representation with a novel coordinate system. In Chapter 1 describe the theoretical frameworks in which this research is situated (Section 1.1), and review relevant theory in graph comprehension (Section 1.2). I describe the methods commonly used in graph comprehension research (Section 1.3), before introducing the graphical formalism which serves as the locus of my case study (See 1.4). I conclude by describing the specific aims of the dissertation and how they are realized in the empirical studies that follow (Section 1.5).

1.1 Representation and the Distributed Cognitive System

Graph comprehension, information visualization more specifically and external representation more broadly are phenomena of interest across a number of disciplines, including psychology, education (especially science and math education), and computer science. These phenomena are studied from a variety of theoretical perspectives with differing goals, from the invention of new formalisms, from scaffolding learning with the formalisms, to developing complex computer systems to afford interaction with the formalisms for the purpose of further discovery. In the following sections I describe the theoretical perspectives that drive my own investigation of these phenomena and inform the design and interpretation of the subsequent empirical studies.

1.1.1 Graphs are External Representations

The power of the unaided mind is highly overrated. Without external aids, memory, thought, and reasoning are all constrained. But human intelligence is highly flexible and adaptive, superb at inventing procedures and objects that overcome its own limits. The real powers come from devising external aids that enhance

cognitive abilities. (...) It is things that make us smart. — Don Norman, *Things That Make Us Smart* 1993, pg.43

The language of representation is slippery and self-referencing. Shown a collection of marks on surfaces, you might label some as art, or pictures, others as diagrams, maps, or schematics, some charts, plots, or graphs, and others also as graphs but you might use air quotes and call them “graph-theory graphs.” Some you will identify as writing, and others, like writing but not; some peculiar or particular system of notation. The linguistic labels you apply to each marking likely depend on your disciplinary background, and are neither exhaustive, nor mutually exclusive. Which of these, are graphs? (Figure 1.1).

What is most critical to the study of how humans use these representations as tools for thinking, is the recognition that they are all instances of a larger class: external representations. Just as psycholinguists are concerned with the psychological and neurobiological factors that enable humans to acquire, use, comprehend, and produce *all* language (not English, or ‘languages using the roman alphabet’, or ‘languages written from left-to-right’), cognitive scientists should be concerned with the factors that enable humans to make use of external representations (not just the ‘graphic’, ‘data-driven’, or ‘computer generated’ variety). In this sense, information designers and engineers of visualization systems have the luxury of specialization. But insofar as we believe that the interaction with these artifacts rely on general-purpose cognitive mechanisms, cognitive scientists do not. To understand how these artifacts function—to study how they are used by humans to construct meaning in support of complex cognitive activities—we must climb up the ladder of abstraction

The term **external representation** came to prominence in the late 1970s, as the new discipline of Cognitive Science emerged from information-processing psychology with a common focus on the existence and nature of mental representation (see Boden, 2006; Lindsay and Norman, 1972; Neisser, 1967; Palmer and Kimchi, 1986). The complexity of external representation, however, was not immediately appreciated. In his treatise on cognitive representation, Palmer argued that mental representations were, “*exceedingly complex and difficult to study,*” so one might

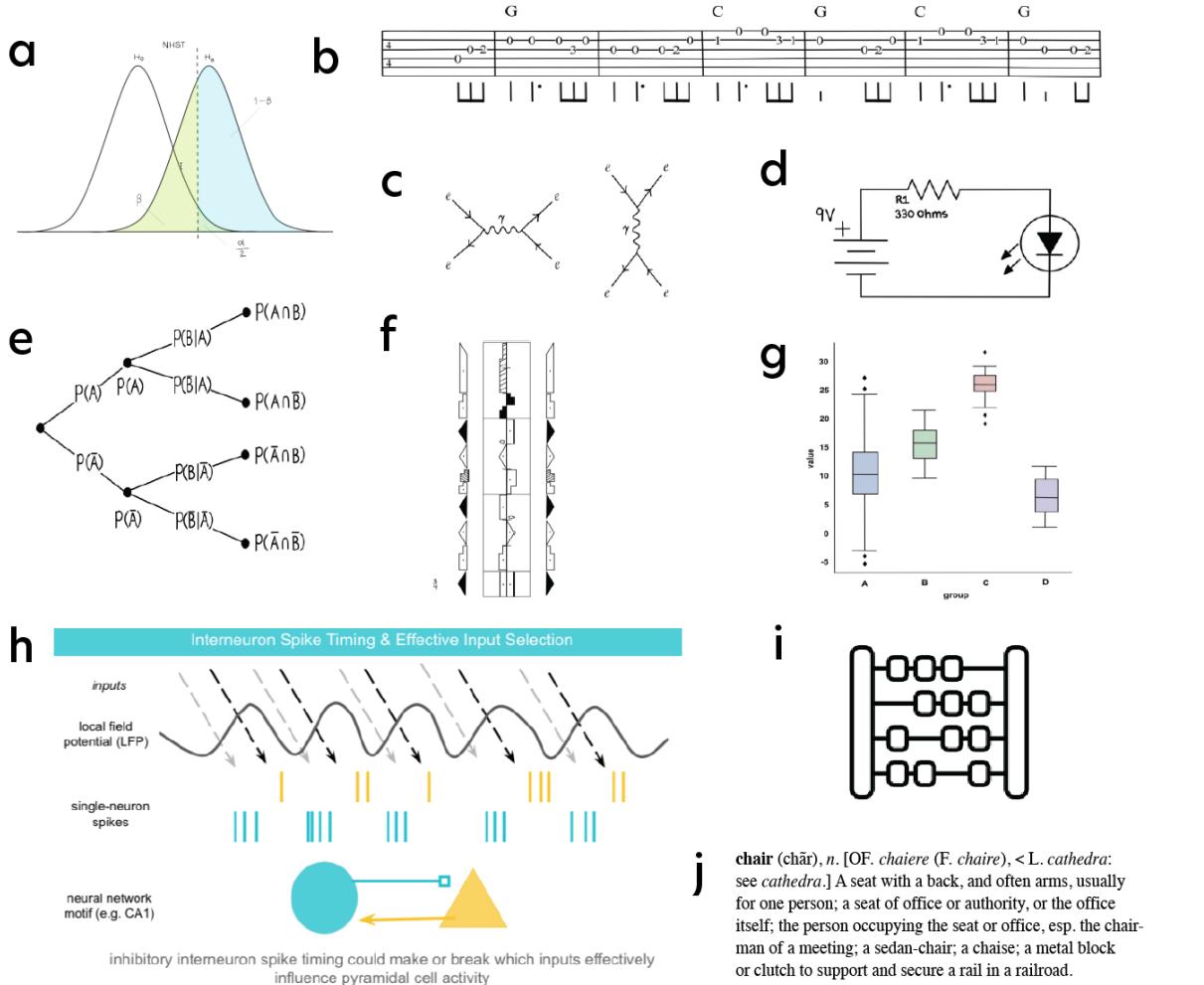


Figure 1.1. A collection of visuospatial external representations. (a) a conceptual diagram indicating key concepts in null hypothesis significance testing; (b) portion of the song 'You Are My Sunshine' in guitar tabs notation; (c) Feynman diagram for an interaction between an electron and anti-electron with exchange of a photon; (d) schematic of a circuit depicting a 9V battery in configuration with a single resistor and LED; (e) tree diagram used in solving Bayesian reasoning problems; (f) Laban notation representing a ballet exercise; (g) boxplot depicting mean, interquartile range and outliers for 4 groups; (h) a figure from a neuroscience presentation that combines multiple representations of related phenomena to orient readers to both the research method and analysis of results; (i) an icon of an abacus—note that the object the icon represents would also be considered an external representation of number; (j) image of the words in a dictionary definition of the word chair (inspired by the conceptual art piece 'One and Three Chairs' by Joseph Kosuth).

start with the examination of “*noncognitive*”¹ representations, as they are “*simple, and easy to study*”² (1978, pg. 262). Subsequent elaboration of representational systems demonstrated there is much to explore with respect to the nature and function of such ‘noncognitive’ structures (see Larkin and Simon, 1987; Scaife and Rogers, 1996). However, empirical work on external representation was often lacking in explicit definition of terms. A study on problem solving with a diagram might refer to the diagram as an external representation and rely on the reader to draw the same antonymic implication as Palmer: an external representation is a representation that is *not* internal. The sensory modality, encoding media, presentation substrate and communicative purpose were left under-specified, allowing the term to serve as a category for *things that can be perceived, that refer to other things*.

A notable exception to this terminological ambiguity was Zhang & Norman who explicitly described external representations as, “*knowledge and structure in the world, as physical symbols (e.g., written symbols, beads of abacuses, etc.) or as external rules, constraints, or relations embedded in physical configurations (e.g., spatial relations of written digits, visual and spatial layouts of diagrams, physical constraints in abacuses, etc)*” (1994, pg.3).

External representations (ERs) can be constructed for any sensory modality and physical substrate, and for a variety of communicative purposes. Information is encoded externally via structures that can be described along a continuum from implicit to explicit, depending on how much effort, or inference, is required in their use (see Kirsh, 1990, 2006). The focus on my work is on those can be seen on some surface. The text on this page is a visual ER, with the letters of the alphabet functioning as symbols referring to sounds that you have learned to assemble into words from which you construct a certain understanding of what I intend to communicate. Similarly, a photograph is an ER, referring via analogy to whatever it depicts. A rich spectrum lies between these symbolic texts (describing the world) and analogous pictures (depicting the

¹Palmer reserves the qualifier *cognitive* for internal representations, designating the external as ‘non-cognitive’. Following a distributed cognitive perspective I characterize *both* as cognitive representations, and prefer the term ‘mental’ to describe those representations not accessible outside the body.

²More “accessible” is perhaps the more generous characterization.

world). I am most focused in what lays between, and in particular, representations that are designed in some capacity to refer to a number of things and their relations by employing space, simplified or schematic forms (and sometimes time). A line graph is a prototypical example, using a series of numbers, lines, points and text arranged on a surface to indicate a quantitative relationship between at least two variables. While a precise taxonomy of the design space of external representations is beyond the scope of this work,³ I focus my attention on one subset of ERs colloquially defined as “graphs” (from the Greek *graphē* ‘writing, drawing’): diagrams that convey relationships between sets of information via visual and spatial variables in a formal coordinate system (see Bertin, 1983; Pinker, 1990). These are not to be confused with another set of ERs colloquially referred to as “graphs”: collections of edges that join in vertices (à la “graph theory”; also referred to as ’node-link’ diagrams). Both varieties of graphs are subsets of the larger class of “diagrams”: external representations that use space and schematic forms to convey relationships between their referents.

The focus of this dissertation is how humans make sense of a further subset of graphs with a particular obscure coordinate system. However, the primarily manipulation—*knowledge of the graphical formalism*—is pertinent to all visuospatial external representations, including those without metric coordinate systems, and the core question —*how humans make sense of representations with which they have no prior experience*—is relevant to external representations in general.

1.1.2 Meaning is Constructed

All meaningful phenomena (including words and images) are signs. To interpret something is to treat it as a sign. All experience is mediated by signs, and communication depends on them. — Chandler, 2017, pg.23

³Though I’ve noted the lack of precision in defining the scope of external representation, there has been no lack of effort in cataloguing (Harris, 1999) and taxonomizing them, from a general theory of symbol systems (Goodman, 1968), to more specific descriptive frameworks for graphical and charting systems (Blackwell & Engelhardt, 2002; Chi, 2000; Parsons & Sedig, 2014a; Shneiderman, 1996; Tory & Moller, 2004) to those concerned with specific domains of data (Aigner, Miksch, Muller, Schumann, & Tominski, 2007; Beck, Burch, Diehl, & Weiskopf, 2016; Blascheck, Kurzhals, Raschke, Burch, Weiskopf, & Ertl, 2017). Two particularly useful (and under-appreciated) are those of Engelhardt (2002) who offers an atomic, generative framework deserving of its characterization as a language of graphics, and Massironi (2001) who offers both a taxonomy and evolutionary timeline.

If external representations are things purposefully constructed to refer to other things, then understanding their referential function falls squarely within the realm of *semiotics*. Semiotics is the study of signs, where a sign is construed as ‘something which stands for something else’—*aliquid stat pro aliquo* (Chandler, 2017). Note this is a larger class of phenomena than external representations which I have (pragmatically) constrained as being purposefully constructed. Signs, conversely, can be naturally occurring: a trail of footprints in the snow, or mud puddles following a heavy rain. The crux of the semiotic puzzle is that to *be* a sign, is to be *interpreted*. Phenomena become signs when meaning is assigned to them. You may have the intuition (and concern) that to implicate semiotics is to open a Pandora’s box where terms like *represent* and *signify* become so complex they risk losing any consistent meaning—and you would be right.⁴ Our task is to introduce the elementary constructs of a particular semiotic approach that can be productively applied to understanding the function of external representations in distributed cognitive systems.

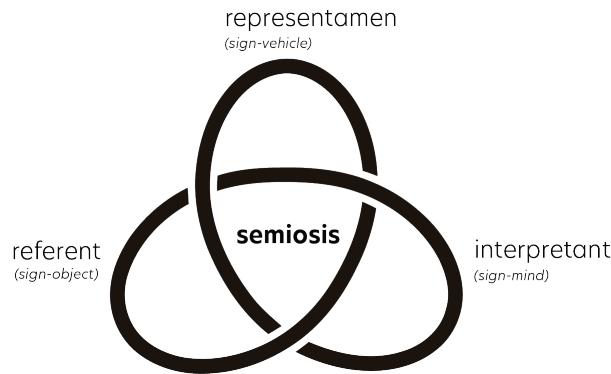


Figure 1.2. The three components of a Peircean sign (referent, representamen, interpretant) are irreducibly triadic.

Imagine you encounter a line graph in a newspaper. Your job as a reader is to develop an understanding (interpretant) of what the graph (the representamen) indicates about some state of the world (the referent). The terms referent, representamen and interpretant are drawn

⁴“No treatment of semiotics can claim to be comprehensive because, in the broadest sense (as a general theory of signs), it embraces the whole field of signification, including ‘life, the universe, and everything’, regardless of whether the signs are goal-directed (or interpreted as being so)” (Chandler, 2017, pg.xvi).

from American philosopher Charles Sanders Peirce, and his general account of the relations that govern representation, reference, and the meaning of signs (Hoopes, 1991). Peirce's basic claim is that a "sign" consists of three parts: (1) an object (referent) that is the thing being signified, (2) an element that signifies (representamen): that which does the referring, and (3) the interpretant: understanding that is made of the referent-representamen relation. Importantly, the entire triadic relation is referred to as a sign or representation, and the dynamics of the relation semiosis or signification. Though Peirce's own terminology changed over the development of his ideas, to avoid confusion we choose here three terms not commonly employed outside of semiotics: *referent* (also: sign-object, or signified), *representamen* (also: sign-vehicle, signifier) and *interpretant* (also: sign-mind, understanding) (see Figure 1.2). The labels we colloquially apply to the material substances that comprise external representations—representation, sign—are in semiotic terms explicitly *not* equated with the material component of the sign. That is to say, the 'representation' is not *the* representation, but only a *part* of it. The sign-relations are *irreducibly* triadic, and while we might for sake of analysis wish to isolate the relation between sign-object and sign-vehicle (for example, how a designer chooses to encode some information), or sign-vehicle and sign-mind (for example, how a reader interprets the encoding), their function is only constituted as a property of all three. This is perhaps more intuitive in psychological terms: constructing meaning is a combination of top-down (knowledge-driven) and bottom-up (stimulus-driven) interpretative processing. To examine how a reader interprets an encoding, we must consider their interaction with the encoding, and prior knowledge of the information being encoded.

Peirce's triadic semiotic is significant to the cognitive science of external representation in two ways. First, it makes explicit the constructive nature of meaning. Peirce's interpretant brings into the signifying function someone or something that does the interpreting: an intelligent process that constructs the translations between signifying elements of the representamen, in order to arrive at some approximation of the referent. In this way, the relation between the 'thing' and the 'representation' is not a direct and determined mapping, but entirely subjective, based

on the interpretation of the observer. Secondly, Peirce's semiosis is dynamic, relying not on the entirety of that which acts as the representamen, but only on the elements relevant in signifying. Later accounts elaborate on subdivisions in the referent and interpretant that pertain to stages of processing in an unfolding chain of meaning (Hoopes, 1991). This aspect has a distinctly cognitive appeal, as it suggests a distribution of meaning-making between the observer and environment; one that occurs via a process in time, not contained solely within artifacts or minds. In the context of cognition, together these features of Peirce's approach are consistent with what we know about the influence of prior knowledge and individual differences in the determination of meaning.

In this dissertation I explore how an individual constructs meaning for a representamen they have never seen before, given extensive prior experience with representations of the same class.

1.1.3 Cognition is Distributed

It does not seem possible to account for the cognitive accomplishments of our species by reference to what is inside our heads alone. One must also consider the cognitive roles of the social and material world. But, how shall we understand the relationships of the social and the material to cognitive processes that take place inside individual human actors? This is the problem that distributed cognition attempts to solve. — Hutchins, 2001, pg.2071

As cognitive scientists, we are concerned not only with the design and efficacy of external representations, but with their mechanisms: how and why they function (or not). These functions are enacted between the artifact(s) and person(s), embodied and situated in their environments and complex social structures. This complexity demands a distributed perspective of cognition, one that extends functions of the mind beyond the individual's skin and skull (see Clark, 1997; Clark and Chalmers, 1998) and distributes them through time and space via material artifacts and members of society (see Hutchins, 2001; Hutchins, 1995). Unlike traditional theories, *distributed cognition* extends the reach of what is considered 'cognitive' beyond the individual to encompass interactions between people and with resources and materials in the environment.

The applicability of a distributed cognitive perspective to research in visualization (Liu, Nersessian, & Stasko, 2007) and human-computer interaction more broadly (Hollan, Hutchins, & Kirsh, 2000) has been successfully argued, and corresponding methods of cognitive ethnography are now widely accepted in VIS and HCI publications. Through cognitive ethnographic techniques (e.g., interviewing, participant observation, in-situ recording) a researcher can determine *what things mean* to the participants in an activity and to document *the means by which* these meanings are created. In this way, cognitive ethnography yields data for exploring cognitive mechanisms, while also feeding distributed cognitive theory by adding to the corpus of observed phenomena the theory should explain.

A distributed perspective on cognition is particularly relevant to the cognitive science of external representations because it not only provides an overarching framework for investigating artifacts and representational processes, but actively encourages integration of ethnographic and experimental approaches. While the study of cognition *in the wild* can answer many kinds of questions about the nature of human cognition in real workplaces, the richness of real-world settings places limits on the power of observational methods. This is where well-motivated experiments are necessary. Having observed phenomena in natural settings, researchers can set about designing more constrained experiments to systematically explore specific aspects of observed situated behaviours. Importantly, distributed cognition does *not* require that every aspect of a cognitive system be examined in every interaction: levels of analysis still apply. But a distributed cognitive perspective does require that the most highly operationalized inquiries of basic processes are contextualized as only parts of a more complex system of factors that taken together, explain behaviour.

In this dissertation I leverage a variety of methods, including observation, experiment, and mouse cursor tracking to explore how an interpreter interacts with their environment to construct meaning for a novel representation. Although I do not directly evaluate the social-situated practices that lead to the construction of prior graph knowledge, it is assumed that these processes are relevant factors that give rise to the graph reading behaviours we directly measure.

1.1.4 Information is Processed

“There is no information without information vehicles. Information vehicles are the carriers of information, the physical material in which the information-for-the-interpreter is encoded.” — Nauta, 1972

In an age of grounded, embodied, and extended cognition, it is rather fashionable to discount information-processing psychology as outdated. However, there is a difference between studying psychological phenomena *as* the processing of information, and studying *the phenomenon* of information processing. The classical conception of information-processing regards the mind as a computational system manipulating symbols to enact representational states. The information-processing psychologist might seek to explain all psychological phenomenon through this lens—behaviour resulting from the propagation of representations, disregarding the influence of the body, modal systems or environment. Contemporary theories that situate cognition beyond the mental are extraordinarily applicable to human interaction with external representations. But so too are some constructs from information processing. In the study of external representations, we are directly concerned with how humans interact with information *via* representations. To the extent that we rely on the notion of information, we cannot escape the notion of its processing: transformation as computation. Importantly, in this work I am not proposing that to adopt an information processing view of visualization requires commitment to a computational theory of mind, nor any strictly sentential/propositional symbol manipulation in the brain. One limitation of early information-processing models of cognition was that they paid ‘scant regard’ to the external world of artifacts and information (see Rogers, 2008). But rather I suggest that by exploring phenomena that require processing of multimedia (i.e. text and graphic) information, we expect that cognitive scientists can improve on these theories by directly addressing the interface between external and internal information, especially in the construction of meaning. Thus throughout this work I will make reference to terms such as information, and knowledge structures in the mind, with no specific commitment to how these are physically realized.

Information Processing Spaces

One particularly productive for considering about the role of information processing in the context of visualization and external representation more broadly is the EDIFICE framework⁵ developed by Sedig & Parsons. As a conceptual model, it provides a structure for thinking about the processing of information (such as goal-directed interaction with a visualization) distributed through the components of a cognitive system. In Figure 1.3 we find five (metaphorical) spaces that together form a *human-information interaction epistemic cycle* (see Sedig and Parsons, 2013, 2016; Sedig, Parsons, and Babanski, 2012).

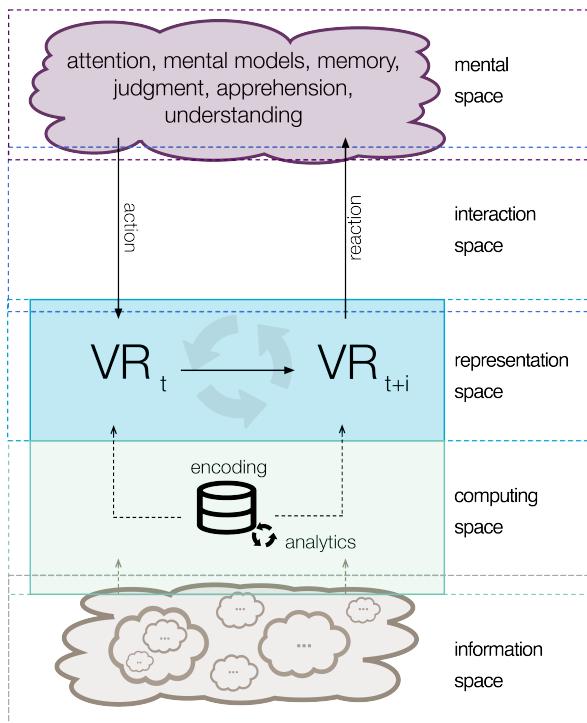


Figure 1.3. The Human-Information Interaction Epistemic Cycle. adapted from original draft with permission of author, Paul Parsons). I emphasize the importance of conceptualizing these spaces as metaphorical, and not simultaneously reifying the layers as physical systems with linear exchanges of information. In practice, information processing emerges dynamically, simultaneously across the material components that constitute the system. This diagram can be construed as a snapshot of this dynamic processing, linearly unfolded in time from left-to-right.

The *information space* consists of the set of information with which users might interact, and the *computing space* its storage and manipulation (i.e. machine computation). In the

⁵Epistemology and Design of human-InFormation Interaction in Cognitive activitiEs

representation space, encoded information is made available for perception. (The ‘space’ of representation is an abstraction, but is reified in computers as ‘the interface’.) The *interaction space* affords exchange of information via action and perception: where the interpreter performs actions and receives reactions. In the *mental space* exists the mind and mental operations that contribute to but importantly do not entirely constitute the construction of knowledge. The model is clearly grounded in the perspectives of information processing and distributed cognition. Though it was conceived in the context of interaction with complex visualization tools, its abstractions can be fruitfully applied to the wider space of multimodal and multimedia external representations. Most importantly, it makes explicit that the design of a visualization tool is a communicative act between designer and user.

The EDIFICE framework offers a productive nomenclature for designating which components of a distributed cognitive system we might be addressing in the context of a particular research project, allowing us to more accurately characterize limitations and desired integrations for future work. For example, a new visualization system that uses machine learning to recommend graph encodings would primarily involve the design of algorithms in the *computing space*, and resultant productions in the *representation space*. A user-study of such a tool would involve measuring the outcome of operations in the *mental space* when an individual interacts with the application (via the *interaction space*). Most importantly, the framework serves as tool for thinking about how the processing of information is distributed across a system of human-visualization interaction: a problem of substantial importance to designers and researchers alike. The authors have applied the framework to describe the relative distribution of information processing across machine and human actors (Parsons & Sedig, 2014b), to characterize the construct of interactivity (Sedig, Parsons, Dittmer, & Haworth, 2014), and as the backbone for a pattern-language to aid conceptualization of novel visualization designs (Sedig & Parsons, 2013, 2016).

In this dissertation I explore the presentation of an unconventional form in the representation space, a reader’s interactions with through the interaction space and iterative

construction of meaning in the mental space in order to develop and understand of to what (information) the representation refers.

1.2 Models and Theories of Graph Comprehension

In this section I review a substantial body of work across disciplines specifically concerned with a smaller subset of external representations referred to as *graphs*, charts or plots. Though these diagrams with formal coordinate systems emerged relatively recently in human history, they have become ubiquitous in the preparation, analysis and communication of quantitative information.

As is often the case with interdisciplinary research, the study of graph comprehension arose from the needs of practice, rather than an invariable march of basic theory. The pioneering graphical inventions of Playfair, Minard and Galton in the 'golden age' of visualization were only made mainstream through inclusion in textbooks (e.g. Brinton, 1914 Brinton, 1914) and standards reports (e.g. "Joint Committee on Standards for Graphic Presentation", 1915), through championing in professional texts (e.g. Tukey, 1977) and essays in scholarly journals (e.g. Cox, 1978; Kruskal, 1975). As the use of such "statistical graphics" spread, guidelines were needed for when and how they could be used to communicate effectively: a call for science to explain the art.

The earliest empirical investigations were published in statistics (Croxton & Stryker, 1927; Eells, 1926; von Huhn, 1927) and consisted of discrete comparisons between bar and pie charts, testing a viewer's performance in judging proportions. Concurrent work in educational psychology (Washburne, 1927) tested secondary school students on their memory of facts learned from bar and line charts, pictographs and tables. Studies of these kind were framed as empirical tests of guidelines offered in textbooks like that of Brinton (1914), but were subject to methodological critiques of construct validity. In contextualizing their results, authors tended to frame outcomes as properties of the representations themselves: *a bar chart is more effective at*

[X] than a pie chart; while contemporary scholars would identify performance as arising from the *interaction between* the individual and representation. This subtle but important difference betrays that the focus of early efforts was on understanding the nature of the representations and their properties.

These type of point-to-point and application-grounded studies would continue for decades, in absence of frameworks, theories or models to guide causal or mechanistic investigation. Work was published in statistics, educational psychology, computer graphics and the burgeoning field of HCI. This would be the case until three developments in the 1980s paved the way for a more coherent, additive body of research to unfold. First, Jaques Bertin's seminal work, *Semiology of Graphics* was translated from French to English by WJ Berg (under the supervision of Howard Wainer) in 1983. Bertin was the first to offer a concise language and structure for decomposing the questions we might ask about what a graphic is, and how it might work. Second, post-cognitive revolution, substantial theories connecting visual perception to higher order cognition had been published in cognitive science —notably Marr (1982) and Ullman (1984). Finally, the ‘mental imagery debate’ was well underway, which saw leading cognitive scientists debating the nature of mental representation. This focus on representation spurred interest in *external* representation, and in particular how graphics are leveraged for problem solving and communication (e.g. Larkin and Simon, 1987).

In the sections that follow, I describe a progression of theoretical development that has shaped the trajectory of graph comprehension research—work that directly addresses the fundamental question: *How are humans able to read graphs?* Our focus will be on the elaboration of general *theory*—accounts of the mechanisms through which our interaction with statistical graphics unfold—rather than individual empirical contributions. We will see examples of theory reasoned from personal experience, appeal to logic, and theory reasoned from experimental evidence. A substantial body of theory has been developed in information visualization and education that addresses the application of visualization and diagrammatic representations more broadly, though (cognitive) theory in graph comprehension can be construed as its foundation,

the backbone of investigations exploring specific phenomena observed within those interactions. Questions like: *What kind of graph is most effective for decision-making?* or *How can we help learners correctly interpret a graph?* rely on general purpose mechanisms of graph comprehension, just as questions of effective linguistic communication rely on the underlying mechanisms of reading and speech comprehension. Figure 1.4 summarizes early theoretical contributions, including a number of general taxonomic grammars and computational efforts that are not discussed in further detail.

The reader will notice that our understanding of graph comprehension did *not* progress via development of *competing* models and theories. Rather, research has unfolded as a progressive elaboration of a vast problem space, with works that shed light on disparate aspects or tasks, and others that expand on prior theory at different levels of detail; iterating rather than refuting. Half of the challenge is deciding what questions need to be answered, and here lies the power and difficulty of such interdisciplinary inquiry.

1.2.1 A Semiology of Graphics — Bertin

“To utilize graphic representation is to relate the visual variables to the components of the information. With its eight independent variables, graphics offers an unlimited choice of constructions for any given information. (...) The basic problem in graphics is thus to choose the most appropriate graphic for representing a given set of information.” — Bertin, 1983, pg.100

Jacques Bertin (1918 - 2010) was a French cartographer, born in the suburbs of Paris and educated in the School of Cartography at the Sorbonne. An esteemed map-maker, he contributed to new methods of cartographic projection as the head of research at France’s National Center for Scientific Research (CNRS) (Palsky, 2019). Yet his most widespread legacy would be the first and most far-reaching effort to provide a theoretical foundation to the design of information graphics, first offered in the 1967 text *Sémioologie Graphique*.

Bertin’s volume resists concise summary⁶, though its most oft-cited concepts, in con-

⁶Any attempt to summarize the 400 page volume would be too brief, and this author is convinced that although widely cited, the depth of Bertin’s intellectual contributions are underestimated on account of opaque linguistic

Early Theoretical Contributions to Graph Comprehension

Year	Author	Key Contributions
1967	Bertin	visual variables; levels of organization
1981/2	Pinker	early version of Pinker 1990, as MIT report
1983	Bertin	english translation by WJ Berg
1984	Cleveland & McGill	ordering of elementary perceptual tasks (codes); <i>(re-articulates Bertin's visual variables with partial accuracy rankings)</i>
1985	Kosslyn	<i>Book review in the J. Amer. Statistics Assoc contained thorough but accessible primer of contemporary information processing psych as applied to graphics</i>
1986	Mackinlay	codification of graphic design criteria in a form that can be used by the presentation tool, including expanded (theoretical) ranking of elementary codes
1987	Cleveland & McGill	expanded set of elementary codes with refined accuracy rankings
1987	Simkin & Hastie	judgement tasks; elementary mental processes <i>(demonstrates interaction of encoding & task; positions Cleveland & McGill in context of Pinker & information procesing)</i>
1989	Kosslyn	analytic scheme for deconstructing graphs; acceptability principles (<i>thorough treatment, framing common graphical intuitions in terms of information processing</i>)
1990	Pinker	first general process account; (<i>schema-theoretic account from information processing perspective</i>)
1993	Lohse	<i>computational (symbolic, GOMS) production-system model predicting scanpath & response time from question & graph</i>
1994	Gillan & Lewis	<i>computational Mixed Arithmetic-Perceptual (MA-P) model derived from task analyses</i>
2002	Shah & Freedman	construction-integration model of graph comprehension <i>(builds upon Pinker 1990 to integrate iteration & prior-knowledge driven processing)</i>
2002/3	Peebles & Cheng	<i>ACT-R/PM based computational model capable of predicting scanpaths on cartesian graphs under questions</i>
2008	Trafton, et. al	<i>argues for explicit inclusion of 'spatial processing' and 'cognitive integration' in existing models</i>

Figure 1.4. Early influential theories, frameworks and models in Graph Comprehension.

temporary writing are the *visual variables* and *levels of organization*, which taken together form a table of perceptual properties: a heuristic for information-visual mapping (Figure 1.5a). Bertin organized the tools at our (external) representational disposal in terms of space (two *planar dimensions*: location on a surface) and the visual (*retinal*) properties along with marks positioned within the space can vary: size, value, texture, color, orientation, and shape. In short, the visual variables offer eight channels into which information can be mapped. Bertin argued these channels have varying capacities for adequately representing different aspects of information: a correspondence between the nature of the information and perceptual requirements for discerning it in graphical form. In an orthogonal scheme, he posited four *levels of organization* that govern what *about* some information we might seek to perceive. Selective perception involves discerning categorical belonging; associated perception grouping like instances; ordered perception discerning step-wise order, and quantitative perception discerning the absolute value of an instance or numeric ratio between instances. Bertin asserted that to map data to a visual variable, the level of organization of the data must correspond to the capacity of the visual variable (Figure 1.5a). Any mismatch is a source of ‘graphic error’ (1983, pg.64).

Bertin envisioned a unifying framework that could govern the design of all kinds of graphics—not only geographic maps or statistical charts. A CNRS colleague reflected that it was the exposure to hundreds of representations from different scientific domains—brought to Bertin for advice—that endowed him with the sort of global perspective required to write a text as comprehensive as *Sémiologie Graphique* (Bonin, 2000). In modern parlance, we would say Bertin offered a structured design space for mapping information to graphical marks. Though it is important to note that these ordered mappings were inferred from a combination of logical reasoning and perceptual experience rather than experimental evidence. Bertin’s treatise is partially descriptive: structuring his observation of the components of graphical communication, and prescriptive: offering guidelines for how and when certain mappings should be made. In

constructions. Bertin also contributed theory on *levels of reading* [pg. 141], *stages of processing*[140], *functions of graphics*[pg.160] and *information processing*[pg. 166]. The motivated reader is strongly encouraged to give ‘Part 1. Semiology of the Graphic Sign-System’ a close reading. (Bertin, 1983)

(A) Bertin (1967, 1983)

LEVEL OF THE VARIABLE			
ASSOCIATIVE (similar)	SELECTIVE (different, groups)	ORDERED (ordered)	QUANTITATIVE (proportional)
VISUAL VARIABLES	Position	Position	Position
	Size	Size	Size
	Color (value)	Color (value)	Color (value)
	Texture	Texture	Texture
	Color (hue)	Color (hue)	
	Orientation	Orientation	
	Shape		

(B) Cleveland & McGill (1984, 1987)

DATA TYPE	
ELEMENTARY PERCEPTUAL TASKS	DATA TYPE
QUANTITATIVE (1984)	QUANTITATIVE (1987)
Position (<i>along a common scale</i>)	Position (<i>along a common scale</i>)
Position (<i>along a non-aligned scale</i>)	Position (<i>along a non-aligned scale</i>)
Length, Direction, Angle	Length
Area	Angles
Volume, Curvature	Slopes*
Shading, Color (saturation)	Areas
	Volumes
	Densities
	Color (saturation)
	Color (hue)

(C) Mackinlay (1986)

DATA TYPE			
ELEMENTARY PERCEPTUAL TASKS	NOMINAL	ORDINAL	QUANTITATIVE
Position	Position	Position	Position
Color (hue)	Density	Length	
Texture	Color (saturation)	Angle	
Connection	Color (hue)	Slope	
Containment	Texture	Area	
Density	Connection	Volume	
Color (saturation)	Containment	Density	
Shape	Length	Color (saturation)	
Length	Angle	Color (hue)	
Angle	Slope		
Slope	Area		
Area	Volume		
Volume			

Figure 1.5. Four contributions ranking perceptual accuracy of visual-spatial encodings.
 Bertin (A) was reasoned from experience, Cleveland & McGill (B) derived from experimental studies with quantitative proportion judgements, which (C) Macklinay (1986) extended for nominal and ordinal data reasoning from existing psychophysics studies, not empirically validated in the context of graph comprehension.

justification of the levels of organization assigned to each variable, Bertin offers a test, a sort of phenomenological self-check (or to the researcher, suggested experimental task) that should convince the reader. In this way, the classification of visual variables can be read as a set of hypotheses for controlled psychophysics experiments. The continued influence of Bertin's work should remind us of the value of the kind *a priori* theorizing required to construct such a theoretical framework. He did not conduct experiments or build models to explain data, but rather imposed a coherent logical structure on a disorganized set of phenomena growing rapidly in importance. Though perceptual experiments would follow, Bertin's visual variables still stand as the most common starting point for information-graphic mapping in visualization design. His work is widely cited in the pioneering research in computer graphics and information visualization, as well as the psychological studies of graphical perception that would begin in earnest in the 1980s.

1.2.2 Elementary Structures in Graphical Perception — From Cleveland & McGill to Simkin & Hastie

We do not pretend that the items on our list are completely distinct tasks; for example, judging angle and direction are clearly related. We do not pretend that our list is exhaustive; for example, color hue and texture (Bertin 1973) are two elementary tasks excluded from the list because they do not have an unambiguous single method of ordering from small to large and thus might be regarded as better for encoding categories rather than real variables. Nevertheless the list ... is a reasonable first try and will lead to some useful results on graph construction.
— Cleveland and McGill, 1984, pg.532

The Semiology of Graphics would not be published in English until 1983, and as graphic displays of information became prevalent in American statistical journals in the early 1970s, calls were made for more systematic inquiry. A “theory of graphical methods” was needed (Cox, 1978, pg.5) in order to overcome the state of “dogmatic and arbitrary” design guidance of the time (Kruskal, 1975, pg.29). William Cleveland and Robert McGill were statisticians at Bell Labs when they answered this call, publishing a series of empirical studies in the Journal of the American Statistical Association (JASA) which they described as theory for the relative

accuracy for a set of *elementary perceptual tasks* readers perform to extract the values of real variables from statistical graphs (1984). In subsequent years, Cleveland & McGill refined their terminology, replacing *perceptual tasks* (1984) with *graphical-perceptual tasks* (1985), *basic graphical judgements* (1986) and finally, *elementary codes* (1987), with influential publications spanning venues of statistics, HCI and popular science. Claims made in their initial 1984 work were tested by additional experiments and deeper engagement with contemporaneous theories of vision, resulting in the much refined 1987 publication ranking accuracy of an expanded set of *elementary codes* (Figure 1.5b).⁷ These codes describe channels available for mapping quantitative information to graphic form. In this sense, the authors re-articulated the visual variables described by Bertin (1967, 1983), and further ordered them according to human accuracy in making quantitative relational judgements. Cleveland & McGill's variables do not match those of Bertin, however, and are admittedly neither exhaustive nor mutually exclusive (1984, pg. 532). One explanation for this discrepancy is their having conceived of the codes on the basis of their personal experience with statistical graphs, while Bertin set out to theorize a structure that could account for the visual-spatial properties of all graphic marks on 2D surfaces.

Cleveland & McGill's approach was partially deductive—structured *a posteriori* from personal experience and perceptual theory (e.g. Stevens, 1975) and inductive, generalizing from reviews of psychophysical experiments (e.g. Baird, 1970), and their own original studies. It is perhaps most accurate to characterize their studies as tests of Bertin's hypotheses for the appropriate visual variables for quantitative perception. The experimental task asked participants—presented with two marked graphic components—to indicate "what percentage the smaller is of the larger" (pg.539), an operationalization of Bertin's test for quantitative perception: "ask the reader the value of the larger sign if a value of one is attributed to the smaller sign" (Bertin, 1983, pg.69). While Bertin reasoned that only the planar dimensions (spatial location) and size can adequately

⁷Nonetheless, the more preliminary 1984 publication remains the most widely cited of their works, with nearly eight times as many citations as the 1987 elaboration [as reported by Google Scholar and Web of Science, January 2021]. This observation reinforces the importance of tracing the intellectual history of theoretical works to find their most mature form, and should serve as a warning against cherry-picking references.

communicate quantitative information, Cleveland & McGill give us the relative accuracy of ten encodings for the same task. Their experimental data support Bertin's hypothesis that spatial location (e.g. position along common scale, position along non-aligned scales) can carry this information most accurately. If *length* is imputed as the size variation of a line (Bertin, 1983, pg.71), and area the size variation of a point, then the data support Bertin's conclusions about the size variable, but not in relation to direction (Bertin's orientation for line) or angle (potentially construed as shape). There is enough discrepancy suggested in the empirical results to warrant further scrutiny of Bertin's criteria for judging a variable as applicable to a particular level, and of the experimental tasks themselves.

Four years later, Northwestern University psychologists David Simkin and Reid Hastie offered JASA a contextualization of Cleveland & McGill's elementary codes, under a framework of information processing psychology (Simkin & Hastie, 1987). Simkin & Hastie emphasized that performance of graphical perception depends not only on the way information is encoded, but also the judgement tasks performed by the human beings for whom the graphs are intended. Building upon Follettie (1986), they differentiated between measurement, discrimination, proportion and comparison judgements (Figure 1.6a). It is important to note that all of Cleveland & McGill's studies used proportion judgements. Follettie, and later Simkin & Hastie, brought awareness to a whole new range of judgement tasks for which statistical graphs are used. Most importantly, they demonstrated that choosing a graphic mapping for a variable of data should not only depend on the data type (Bertin's level of organization) but also the judgement task the designer wants the reader to perform. They offered empirical demonstrations of the interaction between elementary codes and judgement tasks (e.g. comparison judgements were most accurate with simple bar charts (position along common-scale) while proportional judgements were most accurate with simple pie charts (angles)). Moving beyond encoding, they theorized four *elementary mental processes* that could—in an algorithmic sense—explain relative error and response rates across tasks (Figure 1.6b). The elementary mental processes can be construed as visual data extraction steps: ordered in procedures that are executed by the perceptual system in

Elementary Mental Processes

(Simkin & Hastie, 1987)

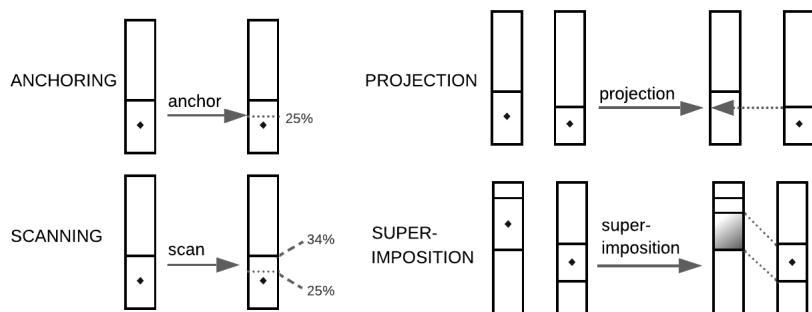


Figure 1.6. Schematic diagram of Simkin & Hastie's theorized Elementary Mental Processes, adapted from (1987)

order to accomplish a judgement task.

Over the course of the 1980s, the use of statistical graphics in publishing and data analysis surged with the development of software packages that made simple visualizations accessible for personal computer users. The cross-fertilization of empirical research between perceptual psychology and statistics demonstrated how demand for design recommendations can drive applied research questions that in turn inspire basic science research. Though the decade began with a focus on mapping information to visual forms, it would end with sophisticated hypotheses about how such mappings would interact with tasks, governed by perceptual rules, to elicit comprehension.

1.2.3 The Rise of Process Theories

Prior to 1980, there had been very little systematic research on the psychology of graph comprehension (Wainer & Thissen, 1981). Over the course of the 1980s, methods and theories from cognitive psychology began to permeate the community in statistics concerned with graphical perception. Simkin & Hastie, notably, were psychologists, though they published their seminal work in the *Journal of the American Statistical Association* (JASA), *not* a journal of applied cognition or perception. Their contribution stood in direct conversation with the earlier work of Cleveland & McGill in the same venue. In 1985, psychologist Stephen Kosslyn

Three stages of Visual Information Processing (as described by Kosslyn, 1989 , with important characteristics noted)

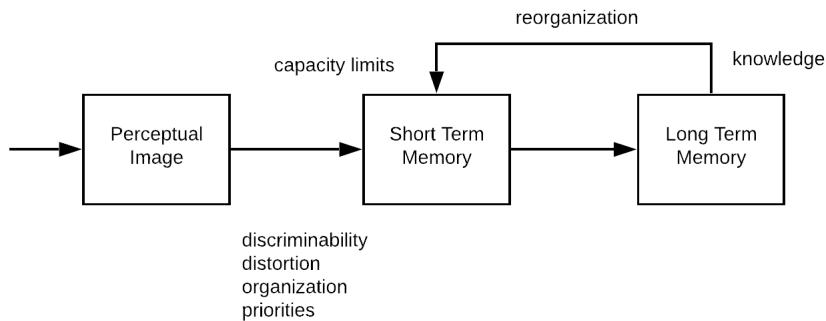


Figure 1.7. A process description of visual information processing, adapted from Kosslyn (1989). The same figure appeared (without linguistic annotation of the important characteristics) in Kosslyn (1985).

published in JASA, a review of five books on charts and graphs, including Bertin (1983), Tufte (1983) and Chambers (1983). Rather than a straightforward critique however, Kosslyn offered a thorough primer on relevant concepts from cognitive psychology contextualized with respect to graph reading. He provided a sketch of contemporary visual information processing (Marr, 1982) and the distinction between short and long-term memory (Anderson & Bower, 1973; Lindsay & Norman, 1972) before addressing the extent to which the practical guidance offered by each book comported with aspects of cognitive theory. Although its citation count pales in comparison to the aforementioned works, the importance of Kosslyn's contribution cannot be overstated. In this cross-disciplinary fertilization, he offered —like Bertin —a structure for thinking about the scope of what questions might be asked of graphical performance. He shared a simple (conceptual, process) model of visual information processing (Figure 1.7) in which graph perception would be situated. To an application-focused community of statisticians *using* graphics, he brought a concise summary of relevant psychological constructs. While previous efforts focused on structural questions of encodings and tasks, Kosslyn drew attention to the way that graph reading unfolds as a *process*.

But Kosslyn's influence would not end there. In 1989 he published an analytic scheme

for deconstructing graphs⁸ into constituent parts, which could then be analyzed at the levels of syntax (configuration of marks), semantics (the meaning that arises from configurations) and pragmatics (conveyance beyond direct interpretation of symbols). This contribution was more structural than procedural, offering a schema for evaluating graphs with respect to acceptability principles reasoned from cognitive theory. But in doing so, he would make reference to a forthcoming publication from his former graduate student Steven Pinker; one that would go on to stand as the most widely-cited theory of graph comprehension.

1.2.3.1 A Theory of Graph Comprehension — Steven Pinker

While experimental psychologist Steven Pinker is most widely recognized for his popular science books on language and human nature, he got his start in the late 1970s as a doctoral student studying visual cognition with Stephen Kosslyn at Harvard. His chapter “A Theory of Graph Comprehension” in the book *Artificial Intelligence and the Future of Testing* would influence research on the design and function of visual-spatial displays across psychology, education and computer science for decades (1990). In fact, the ideas were influential *before* publication, with earlier versions of the theory cited via MIT technical reports from the early 1980s.

Pinker’s theory consists of a series of computational processes that propagate representations of information across components of a theorized human cognitive architecture (Figure 1.8). He proposes that graph interpretation begins with construction of a *visual array*: a relatively raw, minimally processed representation of the information made available to the nervous system via patterns of intensity on the retinas. The visual array is then *encoded* into a *visual description*: a symbolic, structural representation of the scene in a form more efficient for computation with knowledge in memory. A *MATCH process* then compares the visual description with the contents of memory in order to select the correct *graph schema*—a sort of placeholder indicating the structural relation of information for that particular class of graph. Once *instantiated*, informa-

⁸Kosslyn makes a distinction between charts (specifying discrete relations between discrete entities) and graphs (a more constrained form, requiring at least two scales associated via a ‘paired with’ relation).

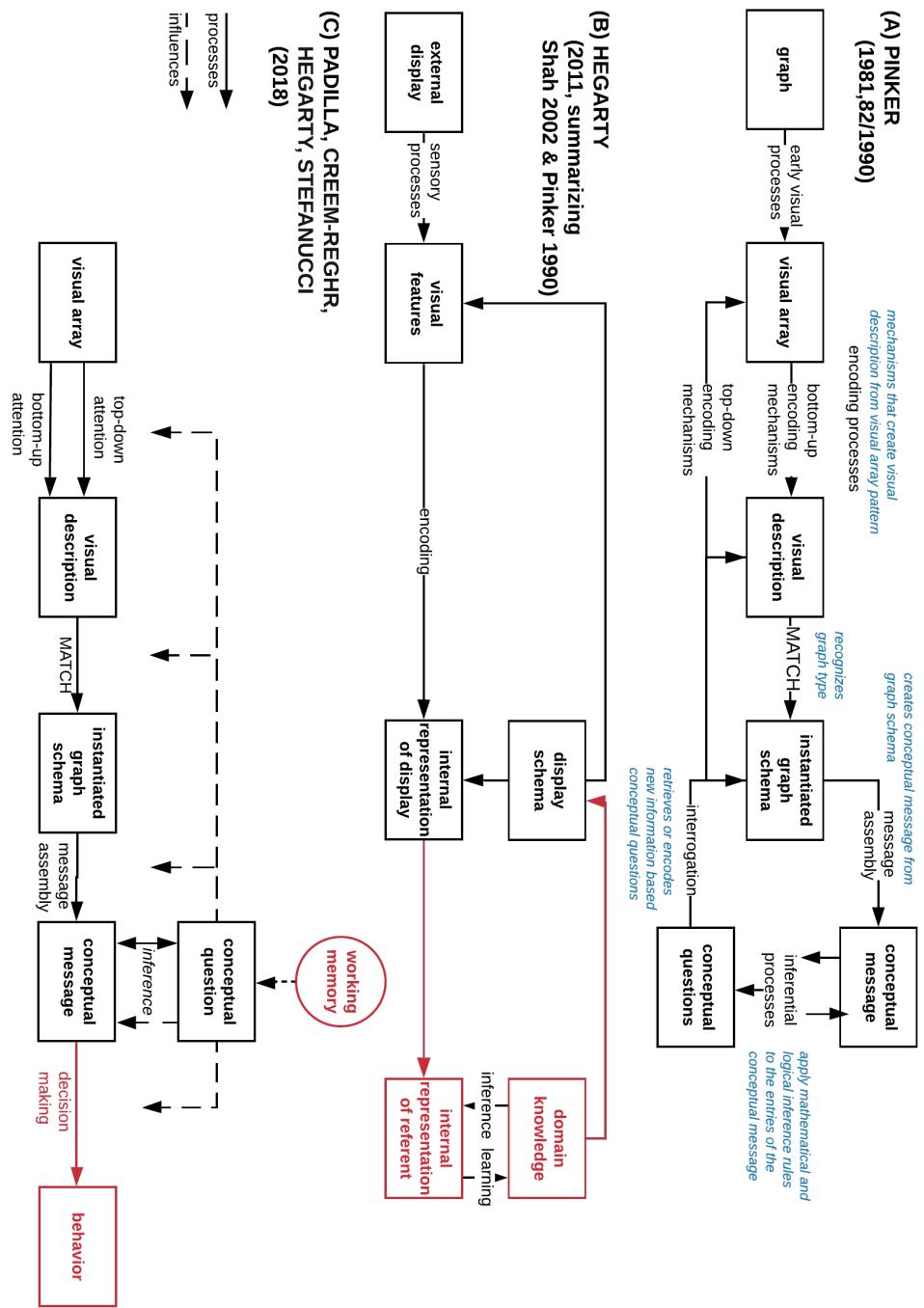


Figure 1.8. Three Information Processing accounts of Graph Comprehension. Italic annotations in blue indicate clarifications, and red indicates changes from prior models. In reading these diagrams, it is important to recognize they represent processes, not components. The boxes in Pinker, for example indicate representations of information, not theorized cognitive structures, like working memory or executive control. The diagrams are not schematics for the structure of a cognitive system, but schematics of how information is processed, and care must be taken to avoid inadvertently reifying them into component structures, which might serve an *implementation* level of analysis.

tion from the visual description is structured according to the relations of the selected schema. By this point, the external representation of the graph has been transformed into an internal representation in some structured, symbolic form that can be interrogated (searched) in order to extract information. Pinker uses the term *conceptual question* to refer to the information the reader wishes to derive from the graph, and *conceptual message* the information that is actually extracted. A *message assembly process* searches the instantiated graph schema for information to translate to the form of the *conceptual message*. But processing capacity limitations prevent all the information from being automatically translated to messages. Rather, the *interrogation* process searches the graph schema for information matching the conceptual question. If it is found, message assembly takes over. But if not, *interrogation* can traverse the prior stages of representation (the visual description, then visual array) until the desired information is found; a top-down search that may require re-encoding the visual array. Finally, Pinker appeals to a general class of (logical, mathematical, and qualitative) *inferential processes* that operate on the conceptual message in service of answering the conceptual question.

Pinker's approach was deeply situated in the tradition of information processing, expressing an orientation toward a computational theory of mind. His explanation functions at Marr's *algorithmic* level of analysis—specifying representations and procedures for transforming them (1982). He offers an exceptionally detailed account of the properties of the representations he proposes (especially the visual description) and how they comport with cognitive theory in vision, memory and attention. The 1990 publication is not an easy read, and it is my personal opinion that its scope is often misunderstood and contribution inadvertently reified as its diagrammatic representation of information processing.⁹ Figure 1.8a is adapted from Pinker's Figures 4.14 and 4.19 which he characterizes as "representing the flow of information specified by the current theory" (1990, pg.104). The diagram depicts the order of representations and names of processes that transform them, but fails to adequately describe re-encoding the visual array (by re-attending

⁹Just as we are drawn to graphs of empirical results, we are drawn to diagrams of theoretical offerings. Readers are warned against assuming that a diagram *entirely represents* a theoretical account, and writers encouraged to explicitly describe the representational role of diagrams in the scope of their theory.

to the graph) or the timecourse of decay of any representation based on the capacity limits of short (i.e. working) memory (e.g. 1990, pg.89). This leads to the misconception that Pinker does not address the role of working memory, or proposes that an entire graph is encoded in a single linear process. Rather, it is more appropriate to construe the diagrammatic representation as a snapshot of the flow of information through a single iteration of a bottom-up (perceptually-driven) loop. We are similarly left wondering "where" in the mind his representations exist. This is not explicitly defined in the process diagram nor the text, but it can be reasonably inferred that all posited internal representations exist in short term (i.e. working) memory, as this is where processing would occur in the context of the cognitive theories he references (with the exception of the uninstantiated graph schema, likely in long-term memory).

Most importantly, justification for the theory rests on a single proposition: that graph comprehension exploits general purpose cognitive and perceptual mechanisms. Pinker's chapter was not the culmination of decades of empirical experimentation with graphs, but rather, the application of contemporaneous theories of vision, memory and attention to the phenomenon of graph comprehension. This statement is not offered in critique, but in observation of the variety of ways that theory is developed. In this case, refutation rests on change to theories of vision, attention and memory, or evidence that graph comprehension is sufficiently different from the phenomena used to construct those theories to warrant special-purpose cognitive mechanisms.

1.2.3.2 A Construction-Integration Model — Shah & Colleagues

An alternative to refuting a theory is refining it, by elaboration (specifying detail) or contextualization (situating in larger scope). In the late 1990s and early 2000s, Priti Shah & colleagues arguably did both: zooming out to describe the iterations of information processing when comprehending a graph, and zooming-in to elaborate the influence of "top-down" factors.

While prior experimental work focused on the perceptual aspects of graph comprehension, Cognitive Psychologist Priti Shah's mid-1990s dissertation work emphasized the role of *cognitive processes* in graph comprehension. Though contemporary Cognitive Science resists a precise

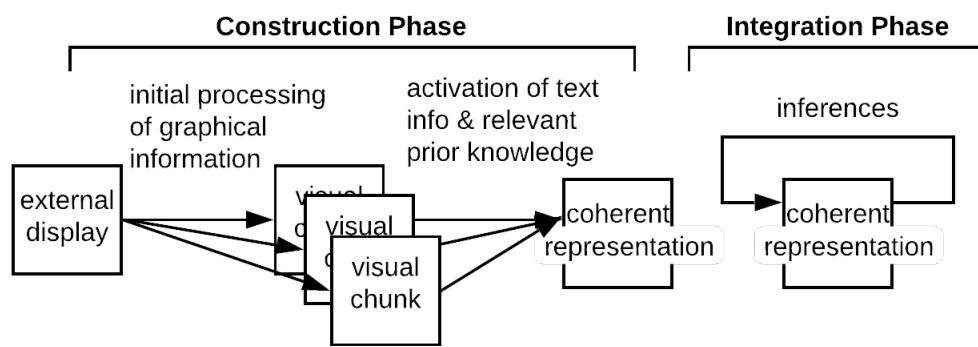
delineation between perception and cognition, in graph comprehension a distinction is typically drawn between sources of information. Perception —information arriving via the senses—is referred to as 'bottom-up' processing, while prior knowledge and computation over internal representations is referred to as 'top-down' processing. Like Pinker, Shah and her colleagues reasoned that graph comprehension would make use of general purpose cognitive processes rather than some special graphics engine in the mind. Drawing inspiration from Walter Kintsch's well-regarded Construction-Integration Theory (1998), Shah elaborated how the processes of constructing meaning with a graph might proceed in the same fashion as constructing meaning from text or linguistic discourse.

Along with Patricia Carpenter, Shah first drew attention to the timecourse of information processing when reading a graph (Carpenter & Shah, 1998; Shah, 1997). Prior perceptual accounts tended to emphasize holistic, pattern recognition processes that allow readers to make the sort of quick proportional judgements used in studies of graphical perception. Carpenter & Shah employed more complex tasks, asking readers to describe graphs and answer comprehension questions. Performance on these tasks, accompanied by measurements of eye fixations, revealed a more iterative procedure was taking place: one that involved a serial identification of visual chunks, followed by inferences and reasoning, repeated until the task goal had been accomplished. Along with evidence of differential task performance based on prior knowledge of semantic content, their studies provided support for the claims that: (1) successful graph interpretation depends not only on appropriate information-to-graphical encoding, but also on prior knowledge and skill of the graph interpreter, and (2) graph comprehension is an iterative, multi-stage process. Publications in 2002 drew more strongly from CI Theory, characterizing the timecourse of processing in terms of two phases: an initial *construction phase*, where visual chunks activate relevant prior knowledge and are integrated into a coherent representation, and an *integration phase*, where inferences are made over the (coherent) representation (Figure 1.9A) (Freedman & Shah, 2002; Shah, 2002). The phases follow in order, though can be repeated, and integration can be followed by further construction, as necessary (Figure 1.9B).

A Construction-Integration Model of Graph Comprehension

(interpreted from Shah 2002)

(A) Two Phases of Graph Comprehension



(B) A Serial, Incremental Process

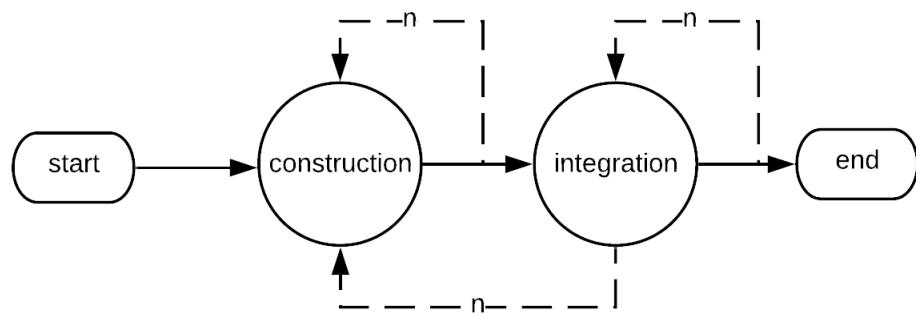


Figure 1.9. A Construction-Integration Model of Graph Comprehension, derived from text description in (Freedman & Shah, 2002; Shah, 2002). Describes two distinct phases of comprehension: the first involves encoding visual chunks, while the second involves higher order cognitive processing over the working internal representation. Figure B describes how integration follows some number (n) of iterations of construction, and repeats some number of times, before processing is either complete, or requires further construction.

The astute reader will ask how Shah's Construction-Integration Model relates to Pinker's (1990) Theory of Graph Comprehension. The answer depends on one's interpretation of each text. In a 2005 review, Shah & colleagues describe their model as differing from Pinker's in that it specifies that prior knowledge (and in turn, expectations) are activated by the encoding of visual chunks, which serve as a top-down constraint on inferential processing (Shah, Freedman, & Vekiri, 2005). Pinker also describes the activation of prior knowledge, though in slightly different terms. Specifically, the MATCH process 'searches' prior knowledge in order to instantiate an appropriate schema (prior knowledge structure) for the type of graph being perceived (Pinker, 1990, pg. 101). In this way, the prior knowledge of graph type is activated by the (symbolic) visual description of the graph (the encoded visual chunk). Since inferential processes act on the instantiated graph schema, this prior knowledge serves to constrain interpretation. What Pinker does not explicitly describe is the activation of prior *domain knowledge*, or any understanding the reader has about the information being represented by the graph, though a generous interpretation would be that he includes this constraining influence under the scope of *inferential processes* (pg. 103), a catch-all term to describe all of the higher order processing (logical, mathematical, judgements and decisions) that one performs *on* the instantiated graph schema. If Shah's *coherent representation* is equated with Pinker's *instantiated graph schema*, then the two accounts are congruous. They are consistent in appealing to general purpose mechanisms, to describing a serial process of encoding, some form of integration with prior knowledge, and inferential processing. They both posit the existence of internal representations: Pinker gives a specific account of a plausible form of these representations, Shah requires only that they exist, leaving the CI model with less explanatory power for mechanisms, but greater robustness to change in the perennial debate on the nature of internal representation. It is *this* author's reading that the these two accounts of graph comprehension are highly compatible, serving to elaborate different aspects of graphical processing at different levels of specificity. While Pinker attends to a computationally-plausible encoding structure for graphical information, Shah attends to the more global timecourse of processing, and iterations of 'perceptual' and 'cognitive' efforts. They

both offer testable predictions about how factors of the graphical display *and* the graph reader should differentially influence task performance.

1.2.4 The Landscape of Contemporary Research

Statistical graphics have never been more prevalent than they are today in scientific inquiry, business operations, or popular media. With such a wealth of applications, it is a good time to be a Visualization Psychologist. But is not easy to *study* the psychology of visualization, because as an applied area of inquiry, both students and scholars alike must navigate an opaque disciplinary milieu. Readers can find relevant empirical research in venues as distinct as journals and conferences of science or math education, learning science, information and library science, cognitive, educational, perceptual or (general) experimental psychology, vision science, cognitive science, and of course computer science—where the conference triad *InfoVIS*, *SciVIS* and *VAST* claim some epistemic authority of the subject matter by virtue of naming rights.

In the two decades since Shah’s Construction-Integration model, we’ve not seen similar overarching, general process accounts of comprehension. Rather, researchers across these fields have progressively elaborated a complex ecosystem of factors that influence performance on graph comprehension tasks. We can organize these factors into three groups: those pertaining to the display, the individual, and the situation.

Display Factors

Research on display characteristics tends to centre on determining the most ideal encoding of information; a question of design. Bertin offered the first experientially-deduced guidelines for mapping data to graphic marks (1967, 1983)¹⁰, some of which were experimentally-tested using relational judgement tasks and ranked by Cleveland & McGill (1984, 1987), and further extended by Mackinlay (1986) who ranked encodings according to theorized perceptual accuracy for communicating quantitative, versus ordered, versus categorical data (see 1.5c). If humans

¹⁰The oft-overlooked footnote to these heuristics is that the rankings are meant to apply when the reader’s *task* is an ‘elementary reading’ (extracting a specific value).

were perceptual computers, this might be the crux of visualization psychology. But we are, of course, more delightfully nuanced creatures. Contemporary research has demonstrated that effectiveness of encodings depends not only on the capacity of a particular type of mark to carry a certain type of information, but also what *about* that information the designer wants the reader to most effortlessly perceive. Ensemble encoding, for example, relies on characteristic performance of the visual system to inform encoding choice when the goal is to facilitate, for example, identification of an outlier, versus recognition of a statistical mean, or apprehension of clusters within the data (Szafir, Haroz, Gleicher, & Franconeri, 2016). Design choices within a particular encoding strategy are nuanced as well, as evidenced by research on the use of color. Color hue has been shown to be particularly effective for encoding data for nominal or absolute value judgements, while color brightness is superior to hue when encoding the same data for *relative* judgements (Breslow, Trafton, & Ratwani, 2009; Merwin & Wickens, 1993). The plot thickens—design choices become more complex—when visualizing more than one variable and the interactions between encoding strategies need be considered. Smart & Szafir recently demonstrated that the shape of a graphic mark significantly influences perception of color and size (2019); whatever the designer’s most informed intentions, their efforts can be thwarted by interactions between decisions they make. Similarly, visual saliency (how ‘attractive’ an area is to the eye) has been shown to influence how humans attend to visual stimuli (Itti & Koch, 2001) though recent efforts to computationally reconcile bottom-up saliency models top-down ‘cognitive’ models have proven ineffective at predicting gaze behavior (Livingston, Matzen, Harrison, Lulushi, Daniel, Dass, Brock, & Decker, 2020). While display characteristics were the focus of the earliest research in graph comprehension, they receive no less attention in modern research efforts. Designers need practical guidance on when and how to use animation (Boucheix & Schneider, 2009; Tversky, Morrison, & Betrancourt, 2002) and 3D (Shah, 2002), how to use signals or instructions to augment a display and scaffold comprehension (Acarturk, Habel, & Cagiltay, 2008; Fox & Hollan, 2018; Kong & Agrawala, 2012; Mautone & Mayer, 2007), and how to most effectively use interaction (Pike, Stasko, Chang, & O’Connell, 2009; Sedig &

Parsons, 2013). Since the time of Cleveland & McGill, research on display characteristics has become increasingly nuanced, revealing more factors that influence how a display should be designed, and the interactions between them.

Individual Factors

Research on individual differences, or factors that give rise to differential performance with the same graphic display, is most common in cognitive and educational psychology and learning science. As Carpenter & Shah argued, "individual differences in graphic knowledge should play as large a role in the comprehension process as does variation in the properties of the graph itself." (1998, pg.97). But what is meant by *graphic knowledge*? In empirical work, graph knowledge is tightly entwined with graph reading abilities and expertise. The terms graphicacy, graphical literacy, graph sense, graphical competence and representational competence are used throughout the literature in psychology and education to refer to a reader's ability to understand (and potentially create) information displayed graphically. If graph comprehension is the act of deriving meaning from a graph, then *graphicacy* is its educational flip side: the ability to perform a graph comprehension task. Some have treated this ability as a foundational step in cognitive development, akin to numeracy and literacy (Friel et al., 2001). Others treat the ability as a practice, implicating the importance of experience and socio-cultural influences (Roth, 2003, 2005). In education in particular, researchers have pursued general learner characteristics that might serve as pre-requisites or predictors of these graphing abilities, including mathematical ability (Curcio, 1987), working memory (Carpenter & Shah, 1998), and spatial reasoning (Velez, M.C., Silver, D., & Tremaine, M., 2005) Ulrich Ludewig's recent doctoral dissertation offers a thorough reconciliation between perspectives of graph comprehension and graphicacy (2018). It is slightly easier to differentiate between ability and knowledge with respect to specific graphs. For example, domain knowledge of the information represented in a particular graph, and knowledge of that particular representation's graphical formalisms. The act of graph reading requires that we use our knowledge of a graph's formalisms to perform some task (e.g. extract

a value, detect a trend) thereby ‘learning’ something about the domain. In my own research I’ve demonstrated that this procedure is not reciprocal. It is much more difficult to use prior knowledge of a domain to ‘reverse engineer’ understanding of a graphical formalism, such as may be required to understand an unfamiliar or unconventional type of graph (Fox & Hollan, 2018; Fox, Hollan, & Walker, 2019). A reader’s understanding of the concepts represented in a graph have been shown to guide not only the reader’s interpretation of the display (Postigo & Pozo, 2004), but early perceptual processing as well (Shah, 2002). In some cases, a reader’s expectations seem to ‘inoculate’ them from true relations presented in the data or lead them to over or underestimate the magnitude of relations. Conversely, domain knowledge has been shown to support comprehension by making readers more likely to ignore ‘noise’ in data (Wright & Murphy, 1984). More recently, Jessica Hullman & colleagues have explored the role of prior beliefs (Hullman, J., Kay, M., Kim, Y., & Shrestha, S., 2018; Kim, Walls, Krafft, & Hullman, 2019) and even judgements of expectations of others (Hullman, Adar, & Shah, 2011) on graph interpretation. Taken together, research on characteristics of individuals has provided strong evidence for ‘top down’ influences on graph comprehension.

Situational Factors

Factors that change comprehension performance of an individual with a particular display depending on the *situation* are the least structured, thus least understood pieces of this factorial puzzle. Affect (emotion) and motivation clearly influence human performance of any task, and although these are characteristics of an individual, we classify them as situational because they are more situationally variable—in the context of a repeated measures study, for example—than the relatively stable¹¹ factors like prior knowledge or ability. *Task* is the most-studied situational factor, though it is at present a hierarchical concept poorly-operationalized across the literature. The term ‘task demand’ is used to indicate a variety of contextual factors, from a relatively low-level step of information extraction (i.e. a micro-step in a larger process, such as identifying a

¹¹Variability, of course, depends on the scope of time under consideration.

location of interest in a graph), to a specific task or goal provided to a reader in an experiment (e.g. extract a value, compare two points, characterize a trend), to the context of some cognitive activity (e.g. analyzing data, making a decision, forecasting, solving a problem), to the communicative intent of the designer (e.g. to inform, educate, entertain, persuade, etc.) In the beginning, there was but a single task: Cleveland & McGill's proportional judgements (1984, 1987). Folettie, followed by Simkin & Hastie elaborated further judgements (measurement, discrimination, and (non-proportional) comparison) (1986, 1987). Bertin also addressed tasks, proposing three "levels of reading" (1983, pg.141). Other tripartite classifications have been proposed in the same vein, all structuring how much of the depicted information the reader need attend to, and how explicit or precise their response should be Bertin, 1967, 1983; Curcio, 1987; Friel, Curcio, and Bright, 2001; Wainer, 1992. In their application of ensemble encoding theories to visualization, Szafir & colleagues offer a parallel taxonomy of four tasks-types that require visual aggregation (2016). These can be partially but not entirely mapped onto the extant tripartite structures. The most complete deconstruction of the concept of task can be found in Brehmer & Munzner's, "Multi-Level Typology of Abstract Visualization Tasks" which surveyed an impressive volume of prior task frameworks in computer graphics and visualization, visual analytics, human-computer interaction, cartography, and information retrieval (2013). A fruitful undertaking for visualization psychology would be to extend this typology to include the tripartite classifications that grew out of education, the lower-level tasks elaborated in vision science, and higher level 'communicative context' that's evident in the structure of the field of visualization itself (Fox, 2020). A strong underlying assumption of most research in graph comprehension (and visualization writ-large) is that the graph designer's goal is to clearly communicate, "the truth" of some data to the reader. Thus, the graph should be maximally informative, and minimally difficult—the graphical equivalent of Grice's maxims for communication. But research in learning science has taught us that sometimes difficulty is *desirable*. Perhaps if my graph is for *learning*, I might encode data differently so as to scaffold a reader's process of discovery and more deeply engage with the data. Alternatively, if the context of my communication is *persuasion* I might use more signals

to direct reader’s attention than I would if the context were exploratory analysis. The role of communicative context is seen structurally through the emergence of specialized workshops at the IEEE VIS conference, but has not yet been systematically investigated across a full range of communicative tasks. My own theoretical intuition—reasoned from design experience and engagement with the literature—is that situational factors are those that present mediating or moderating influences on other individual and display characteristics, at either the time of design, or comprehension.

A primary challenge facing designers and researchers alike is the sheer number of factors found to influence comprehension and the fact that they are typically studied in limited clusters, inconsistently operationalized between studies and across disciplines. This makes it difficult to conceive of the complex interactions that may exist between factors, and how to go about constructing nuanced guidelines for designers. The most comprehensive summaries of factors can be found in (Friel, Curcio, & Bright, 2001; Glazer, 2011; Shah & Hoeffner, 2002) and (Hegarty, 2011) which features a concise set of empirically-grounded principles for display design that would make a useful addition to the wall of any graph designer.

1.3 Methods in Graph Comprehension Research

1.3.1 Task Paradigms and Sources of Data

There are five primary varieties of experimental task used in graph comprehension research. They differ in the way participants are asked to engage with the stimulus representation(s) and offer insights into different aspects of comprehension as a dynamic process. In tasks that require explicit responses, accuracy and latency data are collected. Open-ended responses, including producing descriptions of graphs and making drawings, require resource-intensive content analysis techniques. If performed on a computer, interaction logs (including mouse movements, clicks and keyboard entry) can be gathered, either as a proxy for attention or direct measure of engagement with interactive graph elements. Eye tracking data can also be gathered

for most tasks, shedding light on the allocation of attention across graphical elements. Eye tracking data for graph comprehension are typically analyzed to compare relative fixation time in pre-defined areas of interest (for example, time spent inspecting a graph vs. accompanying text), and the sequence of saccades (scan paths) between graph elements.

1.3.1.1 Perceptual Judgments

Early researchers in statistics and psychology used a perceptual judgement paradigm to investigate the ‘perceptual properties’ of different types of graphs (Cleveland and McGill, 1984; Simkin and Hastie, 1987). Participants were shown two or more graphs with particular data points marked in some way. They were then asked to make quick perceptual judgements as to which of the noted aspects were larger (or longer, smaller, etc.). Speed and accuracy data helped researchers develop guidelines for perceptually appropriate encoding strategies under the presumption that the most important information in a graphic should be easily discriminable. From this research we have guidelines like “data to be compared should be positioned along a common axis” (i.e. grouped vs. stacked bars). The perceptual judgement paradigm is most appropriate for directly comparing graphical forms or encodings on the basis of their perceptual features, but not well-suited for studies involving higher order cognitive processes.

1.3.1.2 Graph Description

Substantial progress has been made in understanding strategies and individual differences in graph comprehension via interview, talk-aloud, and written description tasks. In his anthropological studies, Wolff-Michael Roth presents novice and expert readers with domain-specific graphs found in entry level college texts, asking interviewees to describe everything they can about each graph. After transcribing both verbal and gestural data, Roth analyzes the interviews through lenses of semiotics and activity theory (Roth, 2003; Roth and Bowen, 2003). More concerned with specific interpretations, Carpenter& Shah (Shah and Carpenter, 1995) asked students to describe a series of line graphs and then coded the verbal descriptions according to how the

students characterized each variable (i.e. nominal, ordinal, etc.), and what if any effects were described. Others have coded verbal descriptions for the type of information extracted—discrete comparisons vs. trends (Zacks and Tversky, 1999)—or ‘operation’ described—extraction (qualitative or quantitative), search, reasoning, integration (Ratwani, Trafton, and Boehm-Davis, 2008). Mautone& Mayer took a similar approach in coding descriptions of written descriptions of graphs in their geology classroom intervention (2007). Although content analysis of written and verbal description is time consuming, it has proven a powerful tool in assessing what components of a representation are salient to a participant, as well as the variety of language (often spatial and metaphorical) used to describe the act of representing. Graph descriptions help us understand the strategies underlying how participants use inscriptions to reason. Matching& Verification. In the sentence-graph verification task, participants are presented with a graph alongside a short text and asked to judge if the graph supports the statement in the text. Feeney& colleagues argue this sort of semantic mapping task is a better approximation of goal-directed graph use than more open- ended graph descriptions (Feeney, Hola, Liversedge, Findlay, and Metcalf, 2000). Carpenter& Shah (Carpenter and Shah, 1998) developed a variant of this task, sequentially presenting graphs of either the same or different data in two different formats. They asked participants if the graphs represented the same information, finding that students were often unable to identify informationally equivalent sets. In a similar vein, Strobel& colleagues developed a novel dual- representation task to test whether readers could select the most appropriate graph for a particular task. Informationally-equivalent graphs were presented along with a question, while the researchers measured graph choice, accuracy and response latency to see if participants would prefer the more computationally efficient graph form (Strobel, Sass, Lindner, and Koeller, 2016). They found participants were generally capable of identifying the superior graph. These sorts of graph-message and graph-choice paradigms are most appropriate for probing inferences about the general suitability or message of a graph.

1.3.1.3 Graph Drawing

A number of researchers have asked participants to produce representations in addition to or rather than reading those produced by others. Carpenter& Shah (1995) asked students to reproduce x-y interaction graphs by filling in data points and lines after viewing the data from an alternative perspective. They found students were often unable to reproduce informationally equivalent interaction graphs when the prompt data was displayed with the variables depicted on an alternative axis. In a now classic study, DiSessa& colleagues challenged sixth graders to create their own representations (diagrams) of kinematics (e.g. time, distance, speed and position) A. diSessa, Hammer, and Sherin, 1991). Though more commonly employed in domains that support less constrained representations than graphs (i.e. diagrams in engineering, design, math and physics problem solving) production and drawing is a method of assessing both an individual's representational competency and structural understanding of a representational form; akin to the difference between recall and recognition in memory research.

1.3.1.4 Graph Reading

The most straightforward task paradigm involves asking participants to answer semantic questions using a graph. Sets of questions are designed by a researcher to determine how well the participant is able 'read' the graph, targeting one or more 'levels' of reading typically characterized as either: (1) first-order: extracting a data value, (2) second-order: comparing more than one data value; identifying relations between values and (3) third-order: comparing multiple relations; identifying relations between relations (Wainer, 1992)). Questions are typically posed in multiple-choice form but might also include short answer or free response. This paradigm is frequently used to study the temporal dynamics of multimedia learning and how students integrate information from text and graphics (see Curcio, 1987; Hochpöchler, Schnotz, Rasch, Ullrich, Horz, McElvany, and Baumert, 2013) the order of information processing (see Gillan and Lewis, 1994; G. Lohse, 1993)) and the influence of prior knowledge (see Ratwani, Trafton, and Boehm-Davis, 2008) on comprehension. While it is difficult to study comprehension as a

process using only response accuracy and latency data from an explicit (multiple-choice) task, it has been used in combination with eye tracking in the testing of computational models (Peebles and Cheng, 2003) and to contextualize behavioural data (Strobel, Lindner, Saß, and Köller, 2018). As they are both scalable and adaptable, graph reading questions are the most common approach to assessing graphicacy (graphical literacy) in education.

1.3.1.5 Sources of Data

The explicit measures available via the aforementioned paradigms allow us to assess a reader's accuracy in answering questions (graph reading), as well as the depth and significance of their interpretation (graph drawing, verification, description). Additionally, implicit measures like eye tracking and interaction logging offer a window in the time-course of a reader's cognitive processing.

1.3.1.6 Eye Tracking

The application of eye tracking to graph comprehension was largely inspired by its use in research on reading. In reading and multimedia research, eye tracking is construed as revealing the realtime information processing that occurs when viewing an external stimulus. This application relies upon the assumptions that: (1) eye movements not only represent the distribution of a reader's attention, but also evidence for timecourse of processing ‘in the mind’—the eye-mind hypothesis (Just and Carpenter, 1980), and (2) that the connection between visual attention and cognitive processing occurs immediately, such that as soon as information is perceived, it is available for higher-order processing —the immediacy hypothesis (Rayner, 1998). Eye tracking data is incredibly rich, and care must be taken to determine what signals (and corresponding analytical techniques) are appropriate for addressing any particular research question. In graph comprehension and multimedia learning, the most commonly studied events are fixations (200-300ms period of relative stillness of the eyes), and saccades (quick movements between fixations) (Holmqvist and Andersson, 2017; Scheiter and Eitel, 2017). In many studies,

the visual display is segmented into AOIs (areas of interest) and the relative time and number of fixations is compared between different areas (e.g. Carpenter and Shah, 1998; Peebles and Cheng, 2003). The ordering of transitions between areas of interest has also been studied to test hypotheses about the integration of visual and analytical processing (Ratwani, Trafton, and Boehm-Davis, 2008). More recently, sequence analysis techniques have been applied study differences in the order of processing (Coutrot, Hsiao, and Chan, 2018; Eraslan, Yesilada, and Harper, 2016).

1.3.1.7 Interaction Logging

Interaction logging has long been a foundational data collection technique in Human Computer Interaction and Information Visualization research. If a representation is displayed on a computer, a reader's interactions with input devices can be logged leaving a trace of the reader's attention and engagement with any interactive elements. Like eye tracking, interaction logging offers a rich array of data, depending on what components of the display the researcher chooses to instrument for logging. Log file analysis is non-intrusive and reflects the actual behaviour of computer users.

1.3.2 A Mixed Methods Approach

In this dissertation I leverage a variety of methods to examine the phenomenon of novel graph comprehension. Study 1 includes observation of readers performing an artificial graph reading task in the lab, before a didactic interview and design task. Study 2 includes an artificial graph reading task in the context of a controlled experiment, followed by a graph drawing task. Studies 3 and 4 leverage a graph reading task specially designed to differentiate between alternative interpretations of the graph, with concurrent mouse cursor tracking. In all cases the graph reading tasks utilize a Multiple Choice Multiple Answer item format to maximize available information regarding the nature of participants interpretation of the graph's coordinate system.

1.4 An Unconventional Graph

Most research in graph comprehension involves the kinds of statistical graphics taught in secondary math and science classrooms. In order to explore how adult learners approach new graphical forms, it is most advantageous to start with a formalism that represents information about a domain in which we have reason to expect individuals share sufficient prior knowledge to perform a given task. This allows us to control for differences in behaviour that might arise as a function of individual differences in conceptual knowledge of the domain, versus conceptual or procedural knowledge of the graphical formalism.

1.4.1 A Familiar Domain : Time

On a daily basis we make decisions about how to spend our time. These decisions rely on our facility for reasoning about events in time, their properties and relations. While we are not equally adept at managing our calendars, there is no particularly advanced nor specialized knowledge of time required for one to consider two events and decide which came first, if they overlap, or end at the same time. These types of questions were codified by James Allen into an “algebra” of temporal interval relations (1983). A time interval is defined as a duration (quantity) of time, expressed as an interval between two numbers: a start time, and an end time. In this algebra, Allen defines thirteen atomic relations that describe the possible relationships between two intervals of time (Figure 1.10).

A number of representational systems for reasoning about intervals have been explored in the visual language, mathematics and diagrams literature, due largely to the importance of interval arithmetic in data analysis across the sciences and humanities. But unlike many diagrams designed for learning physics, chemistry, and mathematics, to make sense of time intervals we need only rely on our tacit knowledge from our experience of time. We have selected two informationally equivalent types of time interval graphs, each representing information about events: their start and end time, duration, and the relations between them.

1.4.2 A Familiar Graph: The Linear Model of Interval Relations

When visualized on a two-dimensional surface (such as paper, or computer screen), time and its corresponding intervals is typically conceptualized as a single linear axis —*a timeline*. The sequence of moments along the line is typically consistent with the more general *number line*, but orientation (position in x and y space) is flexible, depending on the nature of the task and constraints of the presentation medium. Humans are capable of remarkable interpretative flexibility in interaction with external representations of time (Coulson and Cánovas, 2009; Fox, de Vries, Lima, and Loker, 2016; Fox and Van Den Berg, 2016; Núñez and Cooperrider, 2013). The most common representation of temporal intervals provides the underlying framework for scheduling artifacts such as Gantt Charts. In the Linear Model of Interval Relations (hereafter LM), intervals of time are depicted as line segments along a one-dimensional timeline which runs from left-to-right on a two-dimensional surface (Figure 1.11 *left*). The left and right boundary points of a line segment indicate their start and end time, respectively, while the length of the segment indicates the duration of the interval. In the LM, the second dimension of the surface (y-axis) is solely exploited to differentiate between intervals, for example, by use of a label. In this way, the y-axis contains no metric information about the interval. As a result, intervals can be sorted along the y-axis in various ways (e.g. in order of start time, by duration, alphabetically by label, etc.). As noted by researchers in visual analytics this polymorphism prohibits the existence of a “universal approach” to visual pattern recognition with the LM, making it ill-suited for applications in exploratory data analysis as well as inspection of extremely large data sets (Qiang, Delafontaine, Versichele, De Maeyer, and Van de Weghe, 2012).

1.4.3 An Unfamiliar Graph: The Triangular Model of Interval Relations

To overcome some of the shortcomings of the Linear Model, alternative representations have been proposed that represent intervals as points in two-dimensional rather than one-dimensional metric space. Notably, Zenon Kulpa developed a series of diagrams to support

For two intervals X and Y ,

each described by a pair of real numbers $[X^-, X^+]$, $[Y^-, Y^+]$

Temporal Relation	Formal Definition
X before Y	$X^- < Y^-$
X meets Y	$X^+ = Y^-$
X overlaps Y	$X^- < Y^- \& X^+ < Y^+ \& X^+ > Y^-$
X starts Y	$X^- = Y^- \& X^+ < Y^+$
X finishes Y	$X^- > Y^- \& X^+ = Y^+$
X during Y	$X^- \geq Y^- \& X^+ < Y^+$
X equals Y	$X^- = Y^- \& X^+ = Y^+$
X contains Y	$X^- < Y^- \& X^+ > Y^+$
X finished-by Y	$X^- < Y^- \& X^+ = Y^+$
X started-by Y	$X^- = Y^- \& X^+ > Y^+$
X overlapped-by Y	$X^- > Y^- \& X^+ > Y^+ \& X^- < Y^+$
X met-by Y	$X^- = Y^+$
X after Y	$X^- > Y^-$

Figure 1.10. Allen's Interval Logic, as developed in (Allen, 1983)

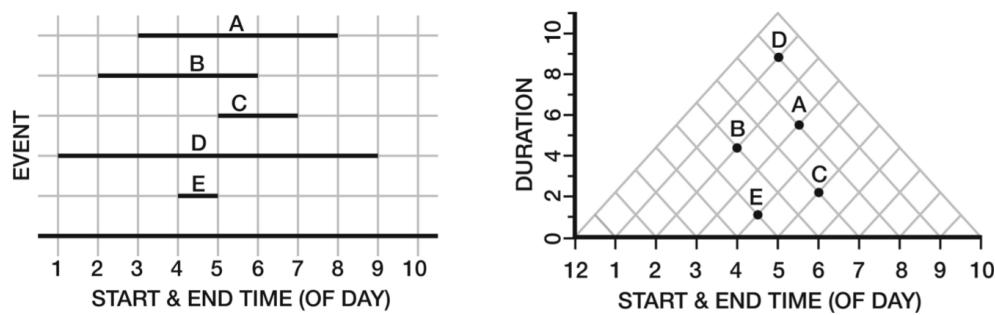


Figure 1.11. Two Graphs for Depicting Interval Relations. Linear Model (*left*) and Triangular Model (*right*)

different aspects of graphical reasoning over general (including non-temporal) interval relations (Kulpa, 1997, 2006). One of these diagrams—termed the ‘MR’ for ‘midpoint-radius’ —employs a novel, triangular coordinate system to explicitly represent both the start, end and duration of an interval. In this way, the second (y-axis) dimension that is not used to display metric properties in the LM is recruited to represent metric duration information in the MR diagram. Kulpa’s MR diagram was applied to the problem of intervals of time by Van de Weghe & colleagues—and and termed the **Triangular Model of Temporal Relations** (hereafter TM) (2007). Their group developed applications of the TM model in fields that require simultaneous visualization of large quantities of temporal intervals, such as archeology and human geography (2012, 2014.)

In the TM time intervals projects the interval relations into 2D space (Figure 1.11 *right*). Each point represents an interval. In the vertical dimension, the height of the point indicates the duration of the interval. The intersection of the point’s triangular projections (using diagonally oriented grid lines) onto the x axis indicates the start and end times of the interval. Under this formalism, every time interval can be represented as a unique point in the 2D interval space, and the characteristics of a time interval are completely expressed by the location of the point. Note that the angle between a point and the x-axis (defined by the gridlines) is an arbitrary constant that can be varied to meet the constraints of the presentation medium (i.e. size and orientation of the page or screen (Qiang, Valcke, De Maeyer, and Van de Weghe, 2014).

1.4.4 On Informational and Computational Equivalence

Following Palmer’s (1978) conceptualization the LM and TM can be considered informationally equivalent because they model the same relations. By Larkin & Simon (1987) all of the information that can be inferred from the LM can also be inferred from the TM, and vice versa. They are not, however, computationally equivalent, as there are some inferences than can be more quickly and easily in the TM—specifically, the *duration* of a given interval. In the TM, the duration can be read off the y-axis via the horizontal intercept of a point; in the LM it must be computed via the difference between an interval’s start and end point intercepts with the x-axis.

Although computational efficiency must be evaluated in the context of the operator performing the computations, if we presume an ideal operator tasked with extracting information using the fewest and simplest operations possible, we can agree the ideal operator would more quickly derive the duration of a given interval using the TM than the LM. Figure 1.12 depicts the full set of Allen interval relations and relevant computations for each graph.

A brief inspection of the TM by even the most experienced graph readers demonstrates its relative obscurity. Given this novelty, can we expect its theoretical efficiencies to be realized by human operators? Qiang and colleagues (2014) evaluated the relative performance of the LM and TM with human subjects. After 20 minutes of video training, students answered data analysis questions with each type of graph. Participants performed significantly better (by accuracy and time) with the TM model on the majority of relation- questions (e.g. before, during, started-by), and only performed better with the LM on property-questions identifying start and end times. These results suggest it is possible for the TM to “outperform” the LM, despite its unfamiliarity to readers. In the studies that follow, we will determine that an individual *comprehends* the Triangular Model if they able to correctly extract start time and end time information from the graph. Will we will determine that they have realized its computational efficiency if they are able to do so more quickly or more accurately than with a Linear Model graph.

1.5 Research Goals

Most research in graph comprehension and information visualization has focused on behaviours associated with the kinds of representations that readers are most likely to encounter in the world. This is a sensical allocation of limited resources. But understanding what happens when we interact with new or unconventional formalisms opens up avenues for addressing questions about the development of representational competencies across the lifespan of a human (*how are new representations learned?*) and of our species (*how do representational conventions come to be?*)

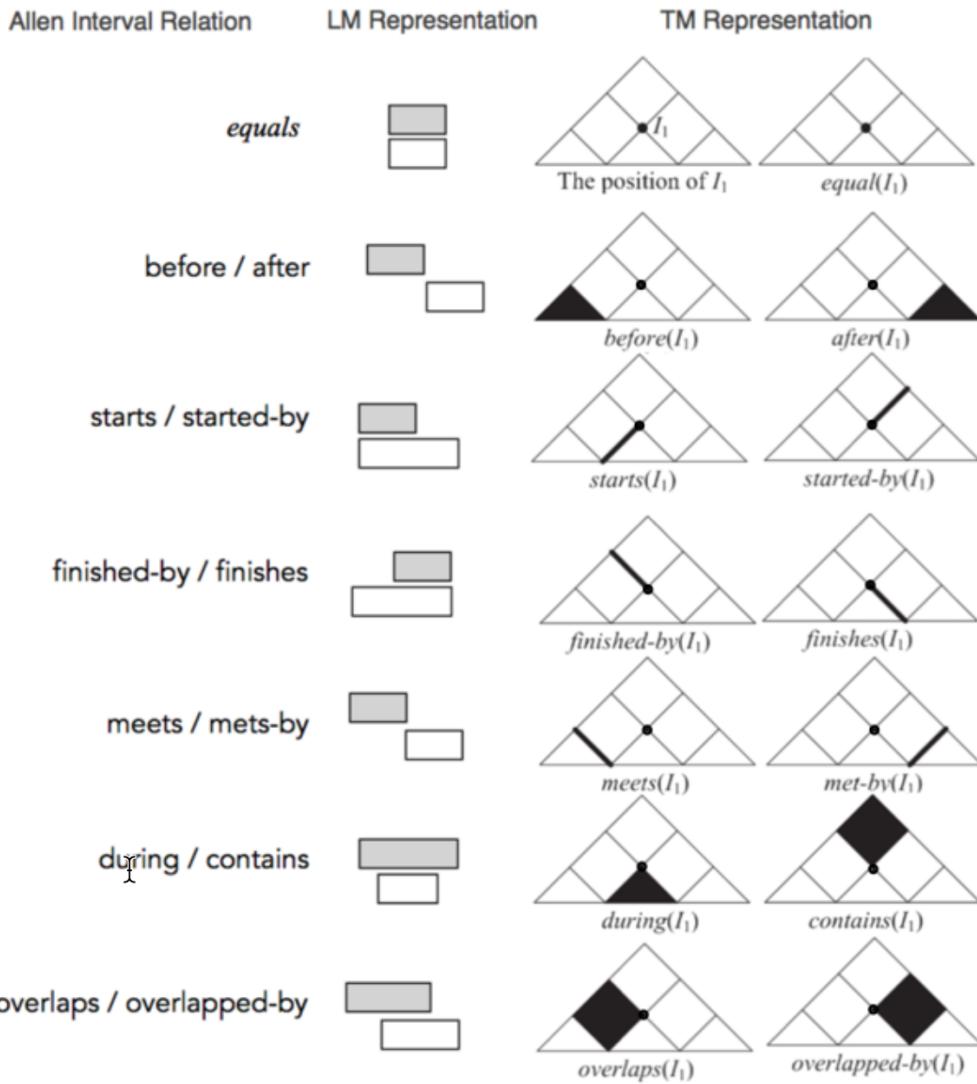


Figure 1.12. Allen (Interval) Relations as represented by LM and TM Graphs. There are 13 relations (12 symmetric and ‘equals’) in Allen’s interval calculus. In this figure, we see the visual-spatial representation of each relation (listed on the far left) first using the formalism of the Linear Model (centre column, intervals as bars), and the Triangular Model (far right, intervals as points). In the TM representation, the area shaded in black indicates the geometric region in which a data point can be found that meets the criteria for the described relation

Answering these questions is beyond the scope of a single dissertation. Rather, in this work I offer a case study in the context of graph comprehension, describing the behaviours that emerge when highly-educated humans with all the requisite perceptual and conceptual apparatuses endeavour to read a simple but unconventional graph. For clarity, throughout the

dissertation I will differentiate between four distinct activities:

- **inventing** a graph: developing a new graphical formalism; as Kulpa *invented* the Triangular Model of Interval Relations in (Kulpa, 1997, 2001).
- **discovering** a graph: determining the rules of a graphical formalism; as in a reader encountering an unfamiliar graph.
- **comprehending** a graph: using the rules of a formalism to accurately extract information; as in a reader performing a task with a familiar graph.
- **designing** a graph: making design decisions within the constraints of a particular formalism; as in a designer creating an instance of a graph for some communicative purpose.

In the studies that follow I address the following research questions:

1. **Study 1:** How do individuals approach interaction with a novel graph? What specific behaviours do they perform and what interpretations of the graph do these produce?
2. **Study 2:** Can errors in interpretation be prevented *via scaffolding*, by providing explicit guidance on the structure of the graphical formalism? Does being asked to use the formalism to represent (rather than read) information change interpretation?
3. **Studies 3A-C:** Can errors in interpretation be prevented *via impasse*, by providing implicit obstacles to incorrect interpretations? Further, does working memory capacity help explain individual differences in interpretation behaviour?
4. **Studies 4A-D:** Can errors in interpretation be prevented *via design*, by altering the visuospatial properties of the gridlines, marks, axes, or orientation of the graph, ostensibly leading readers to instantiate a different (more appropriate) graph schema?

The results of these studies indicate that **discovering the formalism an unconventional graph is much harder than we expect** and that performance is characterized by a combination

of **systematic errors** and **individual differences**. In situating these results I argue that existing theories of graph comprehension are inadequate to explain the how we interpret (and misinterpret) novel graphs. Extensions to process theories and elaboration of their constructs (particularly the graph schema) are needed, and integration of constructs from seemingly disparate areas of cognitive research (such as problem solving, and conceptual integration) may be required.

Acknowledgements

This chapter, in part, includes material accepted for publication as: (1) Fox and Hollan (2023) *Visualization Psychology: Foundations for an Interdisciplinary Research Programme*. In *Visualization Psychology*, Springer; and (2) Fox (2023) *Theories and Models of Graph Comprehension*. In *Visualization Psychology*, Springer. The dissertation author was the primary investigator and author of these chapters.

Chapter 2

Explorations of Explicit Scaffolding

How is it that we develop understanding of new graphical formalisms, when even familiar systems (like scatterplots and line graphs) can prove challenging to interpret (Roth, 2003; Shah and Hoeffner, 2002)? In this chapter, we build upon research on reading and graph comprehension to explore how readers make sense of a simple graph with a novel coordinate system. After generating hypotheses for instructional scaffolding techniques through observation (Study One), we evaluate their efficacy in the laboratory (Study Two). We find that in the face of an unknown graphical formalism, readers are willing to violate several graph-reading norms. It seems that even with explicit (text or image-based) instructions, the influence of prior knowledge from conventional graphs is difficult to overcome. Our results imply that when presenting novel graphical forms, rather than simply telling the reader how it works, the most effective scaffolding techniques will direct readers' attention to the most salient *differences* between their expectations and reality of the formalism. Designers must not take for granted that readers will even notice they are dealing with an unconventional graph.

2.1 Cognitive Aids for Graph Comprehension

Owing largely to their importance in STEM education, techniques for supporting graph comprehension have been a focus of research in the cognitive, computer and learning sciences alike. The most minimal interventions have involved graphical cues—visual elements that guide

attention, akin to gesture and pointing in conversation. Acartürk (2014) investigated the use of point markers, lines, and arrows on bar charts and line graphs, finding that they influenced the way readers interpreted the message of the graphical content. Specifically, they analyzed a combination of eye tracking (relative duration of different Areas Of Interest) and verbal protocol data from participants viewing line and bar graphs with point marker and arrow cues, finding that graph inspection time was shorter for cued graphs, and verbal descriptions were more specific, consistent with the conjecture that graphical cues serve to direct attention. Conversely, uncued graphs were inspected for a longer duration and produced more global descriptions of content. The data suggest that graphical cues can be effective in directing a reader's attention to areas or "messages" of a graph that a designer wishes to emphasize.

In the Information Visualization literature, Kong & Agrawala (2012) proposed the term "graphical overlays" to refer to elements added onto graph content to facilitate specific graph-reading tasks. Reviewing a corpus of statistical graphs in popular media they identified five common types of overlays: (1) reference structures (such as gridlines) (2) highlights, (3) redundant encodings (such as data value labels), (4) summary statistics and (5) annotations, each aimed at reducing cognitive load for particular graph-reading tasks.

In Educational Psychology, Mautone & Mayer (2007) applied techniques from reading comprehension to support graph comprehension in a university geology classroom. In a series of experiments, they presented learners with scatterplot and line graphs augmented by signalling (animations to reveal components of a graph, adding cues to highlight the relationship of depicted variables), concrete graphic organizers (diagrams & photographs of the real-world referents of variables in a graph) and structural graphic organizers (diagrams depicting a relationship analogous to the one represented in a graph). In signalling, cues are added that make the structure of presented information more salient without adding more information. (In this view, information is considered as only what is explicitly communicated. From an information-theoretic perspective, the addition of such signals may in fact constitute additional information.) For text passages, this might include visual indicators such as highlights and underlines, and

propositional indicators such as section headings. To extend this notion to graphs, Mautone & Mayer (2007) employed a segmenting technique, progressively elaborating components of the graph via animation until it was entirely revealed, adding colour shading and arrows to highlight the relationship of variables more explicitly. In text comprehension, advance organizers such as analogies or diagrams are given to learners prior to reading a passage to activate prior knowledge relevant for comprehension. Applying this concept to graphs, the researchers developed two types of graphic organizers: (1) Concrete graphic organizers consisted of diagrams and photographs of the variables in the graph. For the geologic scatter plots used in the experiment, the concrete graphic organizers were diagrams illustrating relevant geologic processes and photographs of sedimentary particles. These organizers were designed to help connect the learner's prior knowledge with the graph content, thus guiding integrative cognitive processes. (2) Structural graphic organizers were defined as cognitive aids given in advance of a graph that directed attention to the "structural relationships" in the graph, independent of the content. For example, a structural graphic organizer for a text about radar waves might be a diagram comparing radar waves to rubber balls bouncing off objects. The researchers predicted that the different types of cognitive aids would target different types of cognitive processing. Both signalling and structural graphic organizers were expected to facilitate organizing processes: where a learner tries to describe depicted information by mentally "organizing" it into a relational structure. They measured organizing cognitive processes via the presence of relational statements from learners that expressed a functional relationship between variables in the graph. Concrete graphic organizers were expected to facilitate integrative processes: where the learner attempts to explain the relational structure by integrating the new information with prior knowledge. Integrative processes were measured via the presence of causal statements, where learners expressed a causal mechanism underlying a relational statement. In a series of experiments, the researchers presented groups of learners with a series of scatter plot and line graphs in the geology domain. They measured the number of relational and causal statements generated in response to targeted questions, in groups provided with only graphs, or versus different

types of cognitive aids. As predicted, they found that learners presented with graphs with the assistance of cognitive aids generated significantly more relational and causal statements than those presented with graphs alone. They found concrete graphic organizers produced more causal statements, while signalling and structural graphic organizers produced more relational statements independent of time spent reviewing the graphs. Taken together, the results of these studies provide support for Mayer's "cognitive model of graph comprehension", inspired by his previous work on text comprehension. Mayer's model, however, is limited in its explanatory power, as it describes only the relationship between "organizing" and "integrating" cognitive processes, without discussion of how and when these processes occur in the context of human cognitive architecture, or how they might be influenced by factors such as prior knowledge and task demands. There is also a lack of granularity in specifying how to apply the cognitive aids constructed to scaffold text comprehension to the domain of graphics: is the presence of descriptive vs. depictive representations important? Do modality and persistence matter? This leaves an opportunity to more precisely elaborate the structure of different types of cognitive aids for graph comprehension, as well as their relative performance in the context of different information processing tasks.

2.1.1 Scaffolding Discovery of the Graphical Formalism

Across all of the techniques described, from graphic organizers to overlays to graphical cues, the design goal has been to activate (or provide additional) prior knowledge and direct reader attention. However, in each investigation it is assumed that the reader has some familiarity with the type of graph being read. Bar charts, scatterplots, time series and line graphs all rely on the Cartesian coordinate system, serving as a common graphical framework. The goal of Acartürk's (2014) work was to understand how different cues affect the message a reader gleans from the graph; a type of gestalt description. For Mautone and Mayer (2007), the goal was to connect the graph scale to conceptual referents to facilitate domain learning. In this way, the existing literature does not differentiate between prior knowledge of the domain and knowledge

of the graphs. The aids did not instruct readers about the “rules” for their representational system. What happens if we’re tasked with a graph that does not look like anything we’ve encountered before? **Might we need a different type of scaffolding to learn a novel representational system?**

2.1.2 Prior Knowledge and Graphical Sensemaking

Process theories of graph comprehension posit a combination of bottom-up and top-down processing (Pinker, 1990; Shah, Freedman, and Vekiri, 2005). So while the design of marks and use of space in a graph is clearly important, so too is the *prior knowledge* a reader brings to the task. When making sense of a graph, we draw on at least two sources of prior knowledge: our knowledge of the domain, and of the graphical formalism itself (Shah and Hoeffner, 2002). Scarcity from either source will impede comprehension in different ways.

Limiting all prior knowledge.

If presented with an unfamiliar graph, depicting information in an unfamiliar domain, I will be unable use knowledge of one to bootstrap inferences for the other. Consider a novice physics student endeavouring to decode a Feynman diagram: without the requisite understanding of particle physics, they cannot reverse-engineer the formalisms of the diagram. Without these formalisms, they cannot draw inferences about particle physics.

Limited knowledge of the domain.

Alternatively, if presented with a familiar graph depicting data in an unfamiliar domain, I can draw on my knowledge of the graph system to learn something about the graph content. If I know a straight line of best fit represents a linear relationship, I can infer that such a relationship exists between the unfamiliar variables in a scatterplot. It is this situation we aim to optimize in STEM education. When we connect our prior knowledge of graphs to the represented variables, we can use that mapping to draw inferences about the represented processes.

Limited knowledge of the formalism.

We are interested in the reciprocal case: an *unfamiliar* representation depicting information in a *familiar* domain. Importantly, by graphical knowledge we are not referring to knowledge of graphs in general (graphical competency or literacy) but rather knowledge of the rules governing a particular graphical formalism. We reason that existing techniques for scaffolding are insufficient for this case, as the information added to the graphs serve only to strengthen the relationship between the graph-signs and (real-world) referents. This fails to address the learner’s scarcity of knowledge for the representational system. If we cannot perform first order readings—such as extracting a data value —we cannot hope to perform higher-order readings, like inferring trends between variables.

With sufficient domain knowledge, we expect that learners might be able to reverse-engineer the formalisms governing an unconventional graph. We wish to scaffold this process to support self-directed graph reading. As a first step, we leverage an obscure graphical formalism using an unconventional coordinate system so that we might shed light on the graphical framework: the foundation of the graph schema (Pinker, 1990).

2.2 Research Goals

We are interested in what happens when experienced graph readers (STEM undergraduates) encounter an unfamiliar coordinate system. Further, we wish to develop and evaluate a series of instructional scaffolds to support self-directed discovery of the coordinate system. In Study One, we start by observing students using the Triangular Model Graph to solve simple time interval questions, and characterize the behaviours that emerge. We evaluate the accuracy of students readings, and then elicit their design insights for how to make the graph ‘easier to read’. In Study Two, we implement four instructional scaffolds inspired by these observations, and evaluate their efficacy in supporting accurate interpretation of the graph.

2.3 Study 1: Observing Interaction With an Unconventional Graph

What strategies do we employ to make sense of an unconventional graph? In this exploratory study we observed students solving problems with the Triangular Model (TM) graph (Part A). After a short interview, we challenged students to design instructional aids making the graph easier to read (Part B). From these data we generate hypotheses for how we might scaffold comprehension for novel graphical formalisms.

2.3.1 Methods

2.3.1.1 Participants

Twenty-three (70% female 30 % male, 0 % other) English speakers from the experimental-subject pool at UC San Diego ($M(\text{age}) = 20$, $SD(\text{age}) = 1$) participated in person in exchange for course-credit. All students were majors in STEM subjects. Participants were recruited in dyad pairs (9 pairs, $n = 18$) to encourage a naturalistic think-aloud protocol. In cases where one recruit was absent we conducted the session with the individual ($n = 5$), altering the procedure only by encouraging them to think-aloud as though explaining their reasoning to a partner. In total, we conducted 14 observation sessions (*9 dyads, 5 individuals*).

2.3.1.2 Materials and Procedure

The Graph Reading Task

The Interval Graph Reading Task is designed to assess a reader's ability to extract information from an interval graph: properly using the graph to reason about the properties of and relations between the depicted intervals of time. Participants were provided with one sheet of paper containing a Triangular Interval Graph with 15 data points, each depicting an interval of time referred to as an "event". A second piece of paper contained task instructions, and the context of a scheduling scenario, where participants were asked to assume the role of an event planner, scheduling events in conference rooms. The scenario instructions were followed by a

list of 16 questions. The questions were open-ended, prompting participants to identify either an event (i.e. a data point in the graph) *or* a reference time (i.e. the start, end, duration, or midpoint). For example, a question testing the “duration” property might read: For how many hours does event [x] last? Figure 2.1 depicts the materials and spatial setup of the task. Participants were provided with pens, pencils and extra scrap paper. A video camera (recording audio and video) was positioned above the table space. After introducing participants to the task and administering an informed consent, the experimenter turned on the video camera and left the room.

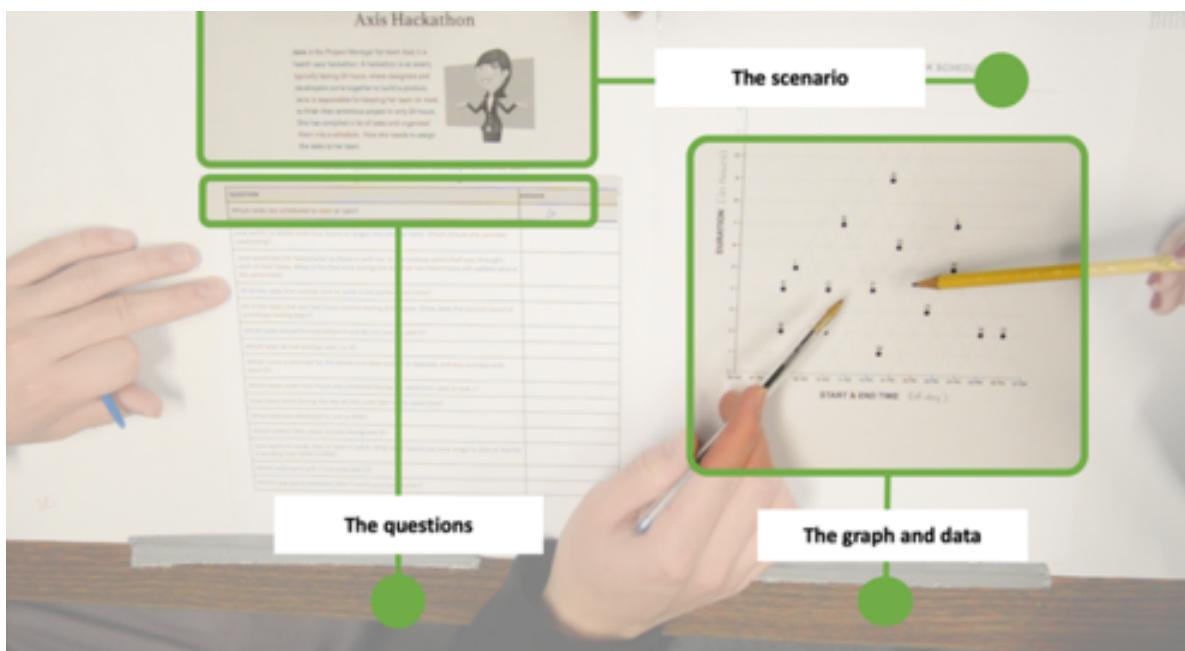


Figure 2.1. Study 1 (Materials) — Layout of the Graph Reading Task. Participants used a TM graph (at right) on one piece of paper to answer questions on a second piece of paper (at left).

Upon task completion, the experimenter conducted a short debriefing interview, prompting participants to explain how they would plot a new data point on the graph. If participants were unable to do so correctly (misinterpreting the graph) we began a didactic interview, prompting students to ask questions they thought might help them discover the rules of the graph system. We responded by only revealing the information explicitly requested, minimizing the effect our teaching might have on the designs produced in the design task to follow. Once students could

correctly plot a new data point (correctly interpreting the graph), we proceeded to The Scaffold Design Task.

The Scaffold Design Task

In the Scaffold Design Task we prompted participants to consider what they could do to make the graph *easier to read* for the next participant. We offered pens and coloured markers and invited them to add instructions and/or visual annotations to the graph. They were free to augment or alter the graph in any way.

2.3.2 Results

2.3.2.1 Graph Reading Task

Overall, observed participants found the task to be very challenging. We evaluated participants' interpretation of the graph by their ability to correctly describe how to plot a new data point (leveraging the triangular coordinate system) in the post-task interview. We found that participants in only 3 of the 14 sessions were able to do so, and their corresponding scores on the graph reading task were high ($M(\text{task-score}) = 12/16$ points, $SD = 1.7$), ($M(\text{time}) = 19$ min, $SD = 30$ s). In the remaining 11 sessions, participants could not (correctly) describe how to plot a new data point. On average, these participants correctly answered only 2 of the 16 questions on the graph reading task ($SD = 2.1$) (Note that the task contained two duration-type questions, for which it is possible to give a correct response, even with an incorrect interpretation of the coordinate system.) What is most notable in these low-scoring sessions was that participants *did persist* in answering *all* questions, and spent about the same amount of time on the task ($M(\text{time}) = 21$ min, $SD = 2$ min). There were no sessions where participants interrupted the experimenter to ask for further instruction or clarification. *So how did these students answer the questions?*

Reviewing the markings participants made on the graphs while performing the task gives us a window into their interpretations (Figure 2.2). Looking first at the lowest scoring sessions (at left), we noticed participants appeared to superimpose the conventional representation for

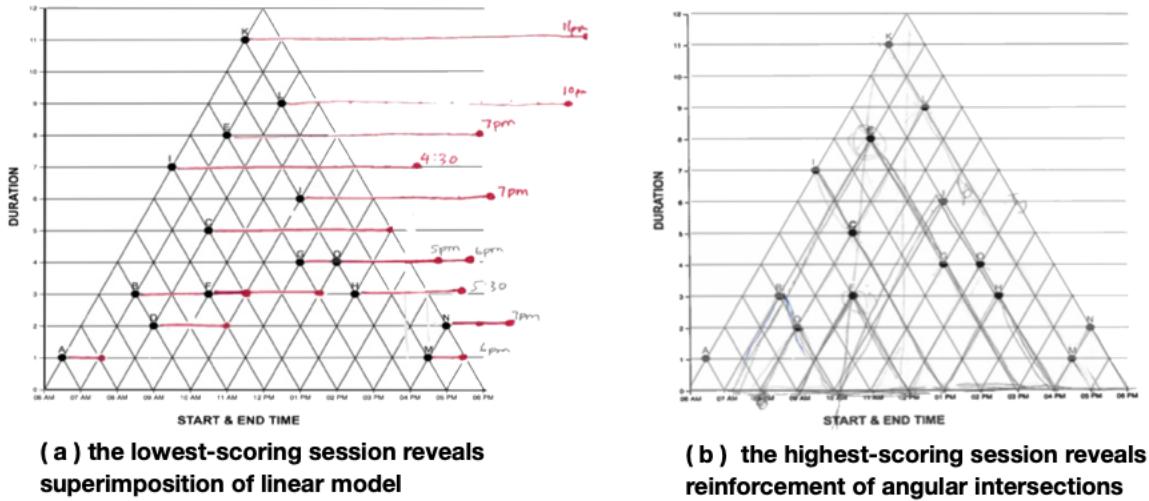


Figure 2.2. Study 1 (Results) — Artifacts of the Graph Reading Task. Participants in the lowest scoring group (*at left*) appear to have superimposed a linear model system atop the Triangular Graph, while the highest scoring participants (*at right*) have reinforced the angular gridlines.

time intervals (the linear model) atop the triangular graph. We refer to this as the *orthogonal interpretation* of the TIG, which relies on participants assuming the data points are situated in a Cartesian coordinate system with a single x and y intercept, where the x-intercept is found by making a single orthogonal projection from the x-axis to the given data point. To hold this interpretation, must also infer that a point represents a *moment* in time, rather than an *interval*, and that the interval is represented by a *line segment* which they must mentally project (or physically draw) atop the graph. They must also decide *which* moment along the interval the point represents: the start time, or end time. In this sense, the *orthogonal interpretation* relies on two kinds of prior knowledge: first of Cartesian coordinates in which a point has a single x-intercept, and secondly of conventions for representing intervals as linear extents, rather than points. This interpretation also requires students to ignore —or assign no meaningful referent to —the graph’s diagonal gridlines. Once constructed, participants could extract information from the *orthogonal interpretation* following the same procedure one would follow for the conventional linear model (LM) graph.

Alternatively, in Figure 2.2 (*right*) we see the artifact from the highest scoring session.

Notably, participants have reinforced the triangular intersections for several points with the x-axis. We *do not* see reinforcement of the intersections with the y-axis, presumably because this is a convention of the coordinate system participants did not need assistance to interpret. Alternatively, In Figure 4-2-right we see the artifact from the highest scoring session. Participants have reinforced the triangular intersections for several points with the x-axis. Noticeably, we do not see reinforcement of the intersections with the y-axis, presumably because this is a convention of the coordinate system participants did not need assistance to interpret. We were curious, however, about the small number of orthogonal projections drawn on the graph. By reviewing the video of this session, we learned that the few vertical lines drawn arose from two situations. In the first instance, the pair was negotiating a possible answer to a 'start' time question. Not finding a data point that directly intersecting the vertical line, they 'veered' off to the next nearest points (see faint line connecting data points F, C, E). The participants subsequently corrected their understanding, and began to leverage the angular gridlines to find the two (correct) diagonal projection to the x-axis. The second instance of orthogonal line drawing occurred when the participants needed to reference a particular time (e.g. "ends by noon"), and (incorrectly) drew a vertical line from the axis instead of a diagonal one. This suggests that even when readers *have correctly decoded* the graphical formalism, they may not necessarily have all the strategies necessary to take advantage of the graph's computational affordances.

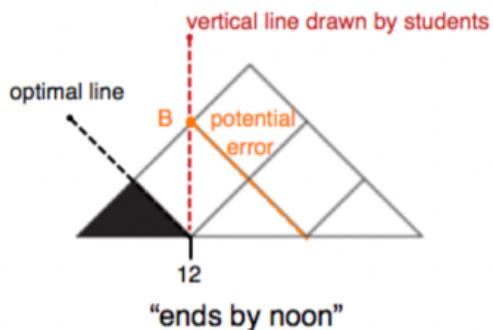


Figure 2.3. Study 1 (Results) — An Orthogonality Bias. When searching for events that end by 12PM, participants erroneously draw orthogonal projection from 12PM, rather than the left ascending diagonal (black shaded region). This gives rise to B as an (incorrect) response.

As illustrated in Figure 2.3 the area shaded in black represents all events that “end by noon”, and the black dotted line the optimal line for demarcating that area. The vertical line is a misstep into Cartesian coordinates that promotes perceptual errors, as in point B (in orange) which lays along this line but ends after noon.

Testing the Orthogonal Interpretation Hypothesis

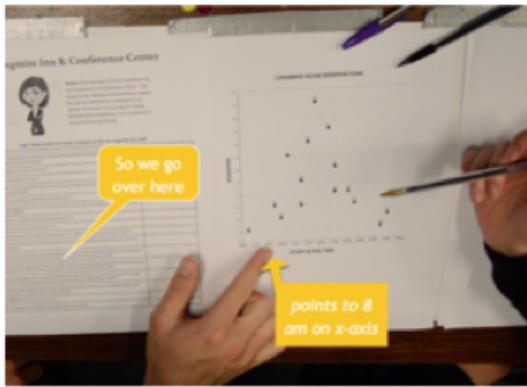
From our review of participants’ graph markings, which were consistent with the procedures they described for plotting a new data point during the interview debrief, we formed the hypothesis that 11 *low-scoring sessions* had formed an *orthogonal interpretation* of the coordinate system, which they subsequently employed to answer the graph reading questions. To test this hypothesis, we constructed an alternative answer key. First, we constructed an *orthogonal interpretation* graph by drawing an orthogonal intersect for each data point to the x-axis and construing this as the start time. We then drew horizontal line segments from each point, with a length determined by the duration given on the y-axis. Using this alternative graph graph, we determined the correct answer for every problem. We then re-scored the task for all participants. Under this alternative answer key, the mean score for the 11 lowest-scoring sessions improved from 2.2 to 8.3 ($SD = 2.7$ points), while the mean score for the 3 highest-scoring sessions decreased 12.3 to 3.0 ($SD = 2.0$ points), supporting the hypothesis that low-scoring participants interpreted the graph in accordance with the conventional linear model, by either drawing or mentally projecting orthogonal intersects from the x-axis.

2.3.2.2 Graph Orienting Behaviour

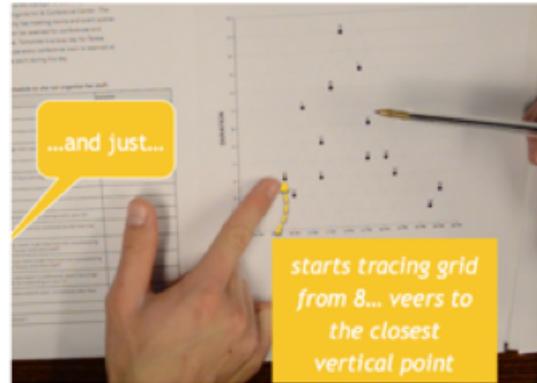
What was the first thing participants did, upon starting the task? When presented with this presumably strange looking graph, how would the dyads approach answering the questions? Would they spend time trying to figure out how the graph worked? Or jump right into answering the questions?

The first question in the problem set addressed the *start time* property, asking participants to identify the events scheduled to begin at a given reference time. In response to this question,

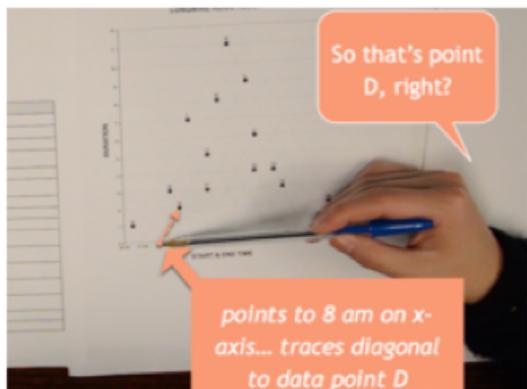
(1) When answering the first question, "Which events start at 8 AM? P1 (in yellow) locates 8AM on the x-axis.



(2) P2 traces up the DIAGONAL gridline from the x-axis, but before reaching the first point (D), they veer to the left, toward the (invisible) orthogonal projection with the reference time.



(3) After some hesitation, P2 (pink) notices P1's error. She points back to 8am on the x-axis, and traces up the diagonal gridline until reaching the first intersecting point : D.



(4) P2 traces the downward diagonal from D to the x-axis, across the axis to the start-time and up the graph to the point, stating that, "this is 2 hours from the start to finish". P1 realizes this forms a triangle.

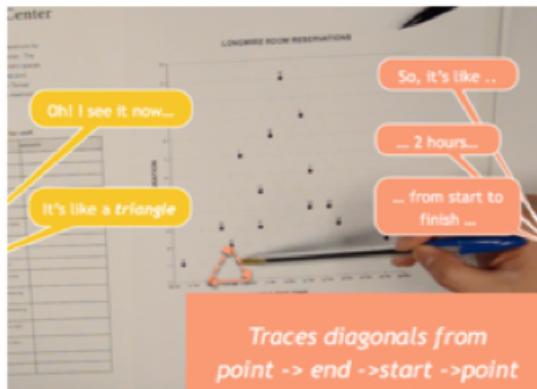


Figure 2.4. Study 1 (Results) — Graph Orienting Behaviour. The highest-scoring groups discover the coordinate system is triangular.

most students located the given start time on the x-axis and traced a vertical line up the page, as one would do if reading a Cartesian scatterplot. **The is the first indication we have in the time-course of problem solving that the students are interpreting the graph relying on their prior knowledge, rather than attending to the novel diagonal gridlines.** Conversely, in two of three sessions where students correctly interpreted the graph, they did so *from the very first question*, either ignoring the orthogonal convention altogether, or “trying it out” before deciding to follow the diagonal grid. (In the third session, participants solved half the questions before

changing their interpretation, at which point they re-solved the previous questions.) In Figure 2.4 we see the series of interactions through which participants in the highest-scoring session came to discover the triangular coordinate system. Across all participants, we observed a number of behaviours that violate conventions of graph reading: (1) accepting that some questions had no answers, (2) needing to add information (extra lines) to the graph to solve the problems, (3) needing to read past the end of the numerical axes, and (4) accepting the presence of information on the graph with no meaning.

2.3.2.3 Debriefing Interview

After completing the Graph Reading Task, we interviewed participants to see if they'd understood the graph. For 11 sessions we had to explain to the participants they'd made a mistake in their interpretation. The most common question students asked was "What do the gridlines mean?" Once we responded they were meant to assist in finding the intersections (emphasis added) with the x-axis, students typically "saw" the triangles. Several students claimed they would not have figured the graph out on their own, but characterized it as a "good" or "clever" representation of the time interval data.

2.3.2.4 Scaffold Design Task

We reviewed the drawings produced in the scaffold design task and grouped them into three clusters based on the primary instructional approach: (1) emphasizing axis intersections (Figure 2.5), (2) annotations/examples (Figure 2.6 *right*) and (3) explicit text instructions (Figure 2.6 *left*).

In Figure 2.5 (*left*) we see examples where participants have drawn attention to the diagonal gridlines and their intersections with the axes by darkening and colouring them. These individuals explained the most challenging part of the graph was realizing they had to look for *two* intersections with the x-axis. In Figure 2.5 (*right*) we see a similar approach; this time the participants have provided annotations to their highlighted intersections. On the left we see a

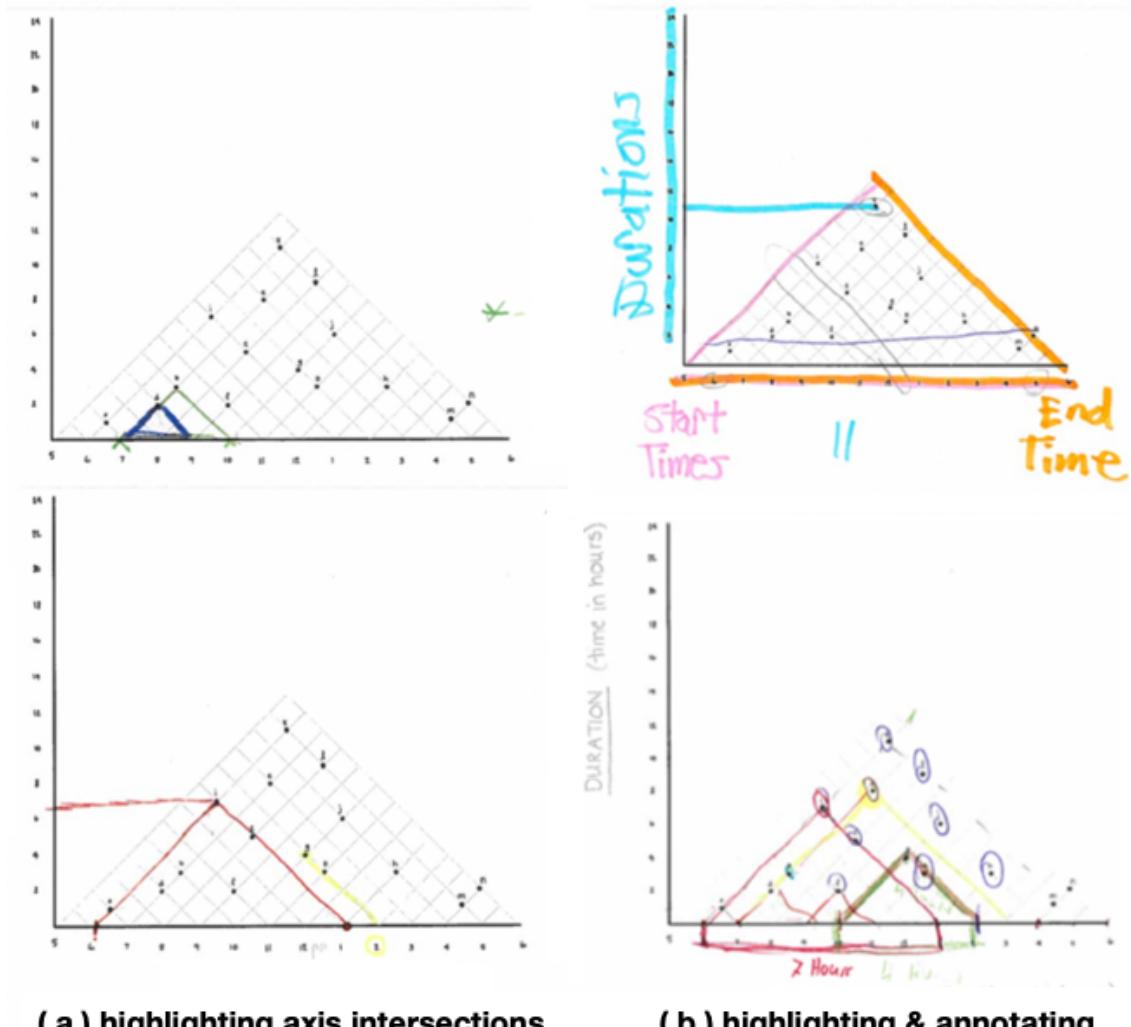
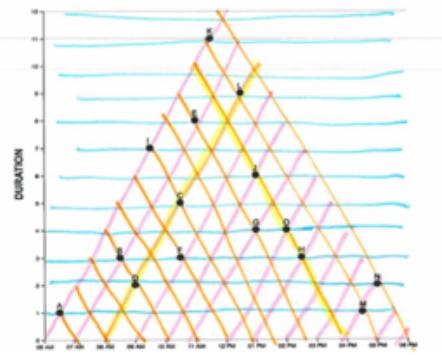


Figure 2.5. Study 1 (Results) — Highlighting Axis Scaffolds. Some participants recommending drawing attention to the dual intersections with the x-axis via highlights.

partial worked example, via the annotation of “7 hours” to the span for the red interval, and on the right, a full specification of the meaning of each axis, in the context of a sample reference interval at the top of the graph. Some learners also suggested differentiating the ascending and descending gridlines with different colour to distinguish “start time” from “end time”.

Finally in Figure 2.6 we find more examples of both explanatory text (at left) and worked examples (at right). Explanatory text included either explicit definitions for the meaning of elements of the graph or procedure for how to extract data from the graph (at right). On the top



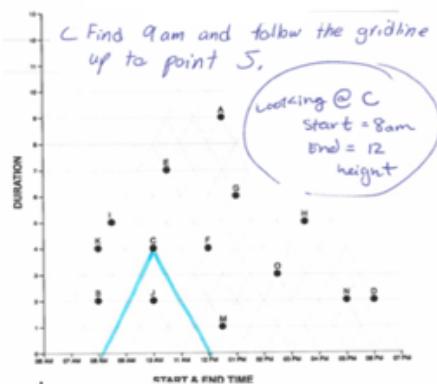
Start Time Each point, B for example, has 3 gridline intersections. one for the start time going diagonal from bottom Left to top Right (7am), the end time going diagonal from bottom right to top left (10am), and a horizontal duration line.

End Time

Duration

the starting point/time for any event (indicated as a dot) begins at the intersection of the left-most line and that event

(a) written instructions



(b) worked examples

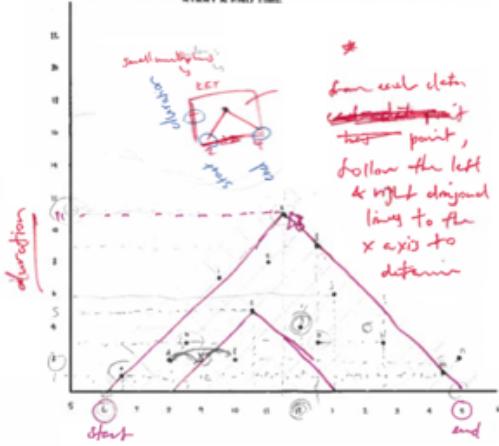


Figure 2.6. Study 1 (Results) — Text and Hybrid Scaffolds. Some participants recommending giving explicit text instructions describing how to find the start/end/duration for a particular data point, or examples of the derivation for a single data point.

right we see a worked example where participants both highlighted the intersection and gave explicit values for a sample point on the plot. Under the graph they added a production rule for finding the start-time of a hypothetical point “S”, indicating that some learners may prefer text instructions. (triangular grid lines were faded in digital scanning)

2.3.3 Study 1 Discussion

The results of Study One suggest the Triangular Interval Graph is challenging for STEM undergraduates. While the graph is elegant in its simplicity—as one participant noted, “*once you see [the triangles], you can’t unsee them*”—most participants failed to independently discover the

rules of the representational system. Rather, they either re-constructed or re-imagined the marks on the page as components of the more conventional representation for intervals. In interpreting this graph students invoked prior knowledge of conventions for graphing in that domain (intervals as line segments) and graphs in general (Cartesian coordinates). When prompted for instructional aids, students believed they could *easily* improve performance of future participants by adding instructions highlighting the multiple intersections of a point with the x-axis. It is important to note that none of the participants recommended redesigning the coordinate system or providing an alternative representation. Despite being told they had struggled for twenty minutes to incorrectly answer a series of questions, they quickly developed an appreciation for the graph's efficiencies. Further, the scaffolds they recommended are substantively different than those explored in previous literature (Acarturk, Habel, and Cagiltay, 2008; Acartürk, 2014; Kong and Agrawala, 2012; Mautone and Mayer, 2007). While some of the recommendations can be characterized as graphical cues, rather than reinforcing the main argument of the graph (e.g. local maxima/minima, salient trend, etc.) their function is to direct attention to the structure of the coordinate system. Both text and image instructions focus on the graphical framework and how to perform a first-order reading, rather than reinforcing the connection between the graph's signifiers and referents.

2.4 Study 2: Testing Explicit Scaffolds for an Unconventional Graph

Inspired by the instructional aids produced by participants in Study One, we designed four scaffolds for self-directed discovery of the coordinate system: two variations of text instructions (positioned adjacent to the graphs) and two illustrations (highlighting x/y intersections). These four designs (along with a no-scaffold control) serve as the randomly-assigned experimental conditions for Study 2. The *conceptual description* (Figure 2.7 *top left*) specifies each component of the formalism and its intended meaning. The *procedural description* (Figure 2.7 *bottom left*)

describes a set of rules for extracting each interval property (start, end, duration). The *static image* (Figure 2.7 *top right*) provides a partially worked example, displaying intersections between a single data point and the x and y axes. The image does not move, and appears in the same position while regardless of the question. Finally, the *interactive-image* (Figure 2.7 *bottom right*), displays the appropriate intersections with the x and y axes when a participant hovers their mouse over any data point.

CONCEPTUAL DESCRIPTION

Describes the graph's signifiers and their intended meaning

A point is an interval of time

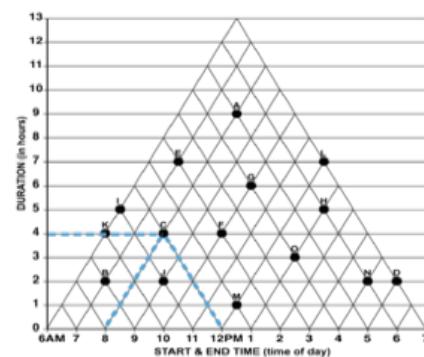
The left intersection with the x-axis along the diagonal gridline is the start time

The right intersection with the x-axis along the diagonal gridline is the end time

The intersection with the y-axis is the duration.

STATIC IMAGE

Depicts x and y axis intersections for a single (example) data point



PROCEDURAL DESCRIPTION

Describes operation for extracting each property of an interval

Start-time: follow the left-most diagonal gridline to the intersection with the x-axis

End-time: follow the right-most diagonal gridline to the intersection with the x-axis

Duration: follow the horizontal gridline to the intersection with the y-axis

Label: the letter directly above the point

INTERACTIVE IMAGE

Depicts x and y axis intersections for each data point on mouse hover

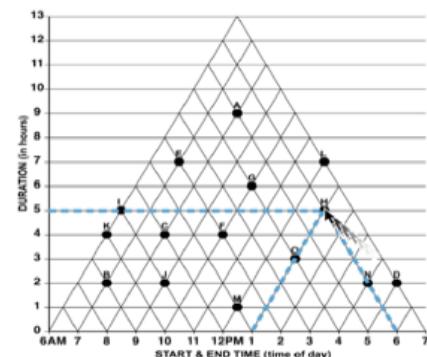


Figure 2.7. Study 2 (Materials) — Scaffold Conditions. We tested the efficacy of two text-based and two image-based scaffolds

In prior work, Qiang and colleagues (2012) demonstrated that the computational efficiency of the TM graph can be achieved by novices after only 20 minutes of interactive video instruction (including feedback). In Study Two we test the effectiveness of our designs by seeking to replicate these results with scaffolding rather than interactive instruction with feedback. Assigning each participant to a scaffold condition, we compare their performance on both the LM and TM graphs, followed by a transfer task testing their ability to draw a TM graph for a small data set.

Specifically, we hypothesize:

- (H1) Scaffolding will *not* affect performance on the LM graph, because readers already understand its conventional coordinate system (i.e. *the utility of scaffolding*).
- (H2) Because the TM graph is unconventional, it will require scaffolding. In its absence, readers will perform significantly better with the LM than the TM (i.e. *the need for scaffolding*).
- (H3) Learners with (any form of) scaffolding will perform better with the TM than LM (replication of Qiang, Delafontaine, Versichele, De Maeyer, and Van de Weghe, 2012) (i.e. *the efficacy of scaffolding*).
- (H4) Learners who solve problems with the LM graph *first* will perform better on the TM (relative to TM-first learners) as their attention will be drawn to the salient differences between the graph types (i.e. *order as scaffold*).

2.4.1 Methods

2.4.1.1 Participants

A total of 316 students at UC San Diego participated (in person) in exchange for course credit (gender: 30 % male, 69 % female, 1 % other; age: 17 - 33 years).

2.4.1.2 Design

The experiment employed a multilevel design structure with 2 fixed factors:

(F1) **explicit scaffold** (between-subjects) @ 5 levels : *none* [control], *conceptual description*, *procedural description*, *static image*, *interactive image* (see Figure 2.7)

(F2) **graph** (within-subjects) @ 2 levels : *linear model* , *triangular model* (see Figure 2.8)

and two random factors :

(R1) **question** (within-subjects) @ 15 levels

(R2) **participant** @ (n = 316) levels

Participants were nested within condition, which was fully crossed with graph. Two additional fixed factors, (Block Order : *LM-first*, *TM-first*) and (Scenario Order: *A-B*, *B-A*) were counterbalanced to facilitate the repeated measures design. Questions were fully crossed with condition, such that each participant was randomly assigned to one explicit scaffold condition, block-order, and scenario-order, in which they completed two blocks of 15 questions, one for each **graph**.

2.4.1.3 Materials

Graph Reading Task

The Interval Graph Reading Task was an extension of the paper-based task used in Study 1, implemented via a computer-based web application. Participants completed two blocks of graph reading questions; one for each interval graph (Triangular Model [TM] and Linear Model [LM]). Each block included 15 multiple choice questions (adapted from the questions used in Study One). The questions were presented one at a time, and participants did not receive feedback as to the accuracy of their response before proceeding.

For the first five questions of each graph reading task, participants saw a scaffold as designated by their randomly assigned explicit scaffold condition. On the following ten questions, the scaffold was not present. Examples of each scaffold for the TM graph are shown in Figure 2.7. Equivalent scaffolds were displayed for the LM graph. The order of the first five (scaffolded) questions was the same for each participant, while the order of the remaining 10 were randomized. For each question, the participant's response accuracy (correct, incorrect) and latency (time

from page-load to “submit” button press) was recorded. Because each participant completed the reading task once with each graph, we developed two matched scenarios: a project manager scheduling tasks (scenario A), and an events manager scheduling reservations (scenario B) (see Figure 2.8). In each scenario, an equivalent question can be identified in the other pertaining to the same interval property/relation. For example, in scenario A the question mapping to the “starts” property reads: “Which tasks are scheduled to start at 1 pm?”, and the correct answer consists of 2 tasks (Fig. 5 – left – tasks O & H). In scenario B, the equivalent question reads: “Which reservations start at 8:00 AM?”, the correct answer referencing 3 events (Fig. 5 – right – events D, C & L). For the LM graphs, intervals were sorted in order of duration, with the longest appearing at the top of the graph. A pilot study on Amazon Mechanical Turk using the LM graph revealed no significant differences in response accuracy or latency between the scenarios. The four graphs constructed for the study are shown in Figure 2.8, and the questions and scenarios are available in Appendix A.1.

Graph Drawing Task

In the graph drawing task participants were prompted to construct a TM graph, and answer two graph reading questions (These responses were not analyzed, but used to record drawing response time). Participants were given a sheet of isometric dot paper and a data set of 10 intervals (see Appendix A). The isometric dot paper is ideal for this task as it equally supports the construction lines at 0, 45 and 90 degrees, thus minimizing any biasing effects of the paper on the type of graph the participants choose to draw. Participants were directed to draw a triangular model graph of the data (“like the triangle graph you saw in the previous task”). They were provided with pencils, erasers and a ruler for the task. Before continuing to the next page, they were asked to verify that they had written a title for the graph, a label for each axis, a label for each tick mark and each data point. Finally, they were asked to answer two questions similar to those presented in the graph reading task but made more difficult by requiring the participant to detect a pattern in the data set and identify an outlier. The task was designed to assess the depth

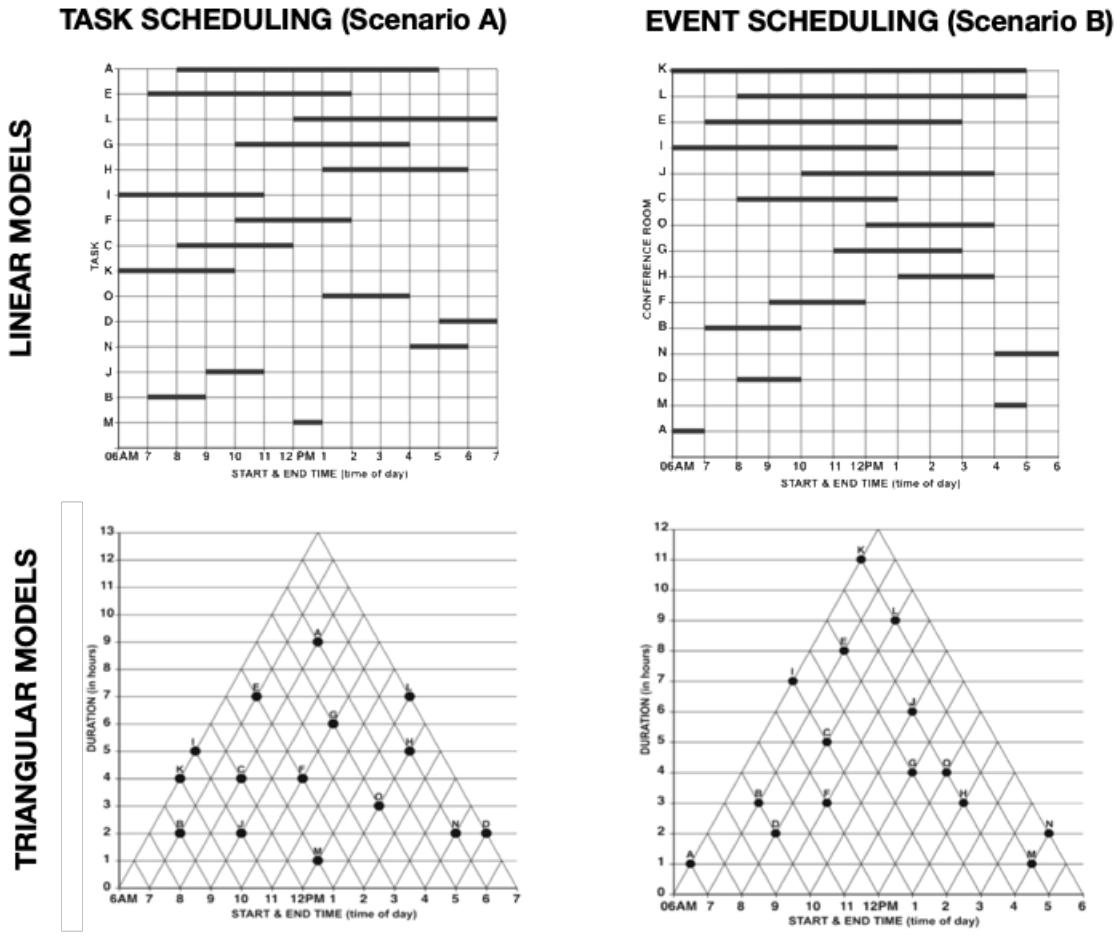


Figure 2.8. Study 2 (Materials) — Stimulus Graphs. Note that each scenario (column) represents the same underlying dataset (schedule of events). The position of the data points differ in the second scenario. Scenarios are counterbalanced across graphs, such that each participant completes the triangular task with one scenario, and the linear task with the other.

of participants' understanding of the representational system of the TM graph. This transfer task allows us to assess each scaffold's efficacy in helping the participant learn the graphical framework of the TM, beyond their ability to read inferences from it. We expect that accurately drawing the graph requires deeper understanding of how the graph works. The materials for the graph drawing task are available in Appendix A.1.

2.4.1.4 Procedure

Participants completed the study individually in a computer lab. They completed the two graph-reading tasks in sequence, one with a TM graph and the other with an LM graph (order counterbalanced). Afterwards, participants completed the graph drawing task. The entire procedure ranged from 22 to 66 minutes.

2.4.2 Results

Figure 2.9 depicts the mean response score by graph, scaffold condition, and graph-order. We see that the mean scores on the linear graph task (green squares) are consistent across both graph order and scaffold condition. In contrast, the triangular graph scores (red triangles) vary by scaffold condition. It appears that accuracy on the triangular graph task only approaches that of the linear graph in the interactive image condition.

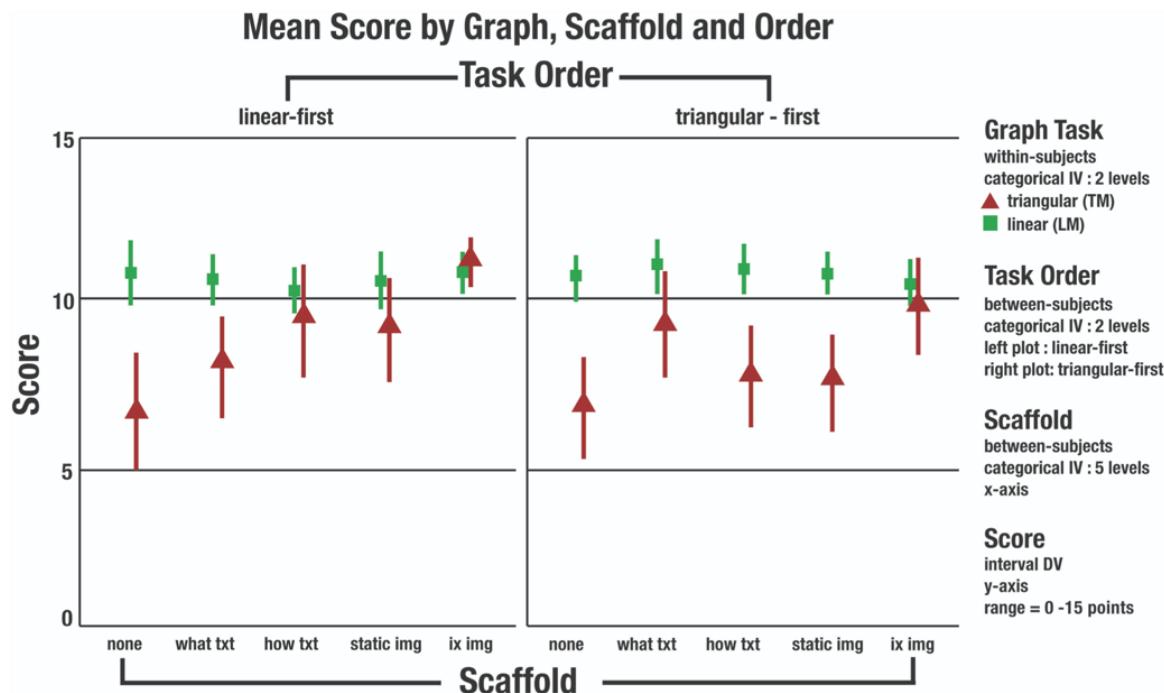


Figure 2.9. Study 2 (Results) — Total Score. Mean response score by Graph, Scaffold Condition and Task Order. LM scores (squares) remain steady across scaffold (x-axis) and graph-order (right/left plot), while TM scores (triangles) differ by scaffold, highest in the interactive image condition. Error bars indicate 95% confidence interval.

2.4.2.1 (H1) The Utility of Scaffolding

Inspection of linear graph scores in Figure 2.9 (green squares) suggests that explicit scaffold condition (x-axis position) did not affect performance on the linear graph task. To test hypothesis that scaffolding is not universally helpful, but rather is only necessary to facilitate discovery of the triangular coordinate system, we compare performance across explicit scaffold conditions on *just* the linear model graph task. A Kruskal-Wallis rank sum test supports this hypothesis, indicating the median linear graph score did not differ by scaffold condition ($\chi(4) = 0.4, p > 0.1$) (see A.1).

2.4.2.2 (H2) The Need for Scaffolding

Inspection of *difference* between linear (green square) and triangular (red triangle) scores in just the *no scaffold* [control] condition (first column of x-axis) of figure 2.9 suggests that in the absence of scaffolding, readers struggled to interpret the triangular graph. To test the hypothesis that scaffolding *is needed* to improve interpretation of the triangular graph, we compare scores on the linear (vs) triangular graphs in only the *no scaffold* [control] condition. Based on our observations in Study 1, we expect that in the absence of scaffolding, scores on the triangular graph block will be substantially lower than the linear graph. A paired Wilcoxon signed rank test (with continuity correction) supports this hypothesis, indicating that accuracy scores in the control condition were higher with the Linear graph than the Triangular graph, a statistically significant and very large difference ($W = 1490.00, p < 0.001; 95\% CI[0.67, 0.88]$) (see A.2).

2.4.2.3 (H3) The Effectiveness of Scaffolding

Inspection of the difference between linear and triangular graph scores *across* scaffold conditions in Figure 2.9 suggest that our second hypothesis is likely not supported: at least some of the scaffolds were not effective in improving triangular graph performance. To test the hypothesis that any form of scaffolding will replicate the findings of Qiang et. al (2012) (significantly better accuracy on TM than LM graph), we calculated a performance increase score

(Triangular Graph score - Linear Accuracy score). Positive values indicate better performance with the TM than LM, with a range from -15 to +15. In all but interactive image condition, the median difference score was less than 0, indicating that only the interactive image yielded more accurate performance on the TM graph than the LM graph. However, the performance increase was very small: a one-sample t-test indicates that the average performance increase in the interactive image condition was not significantly different than zero (no improvement) ($t(67) = -0.6, p = 0.6$).

Inspection of the triangular graph scores (green squares) across scaffold conditions in Figure 2.9 *does* suggest, however, that the type of explicit scaffold has an effect on triangular graph performance, even if it does not improve performance to the extent that would realize the computational efficiency of the TM graph. To quantify the effect of explicit scaffold on TM performance, we fit a mixed logistic regression model on question-level accuracy (correct/incorrect) for the triangular graph task¹. We included random intercepts for participants and questions, and fixed effects for explicit scaffold condition, graph-order, and scenario-order. A likelihood ratio test comparing a model with main effects to a second model including the interaction between fixed factors indicates that the interaction term does not improve model fit ($\chi^2(22, 9) = 11.87, p = 0.539$). The explanatory power of the final model is moderate (conditional $R^2 = 0.47$) with the part related to fixed effects explaining 9% of variance.

Wald Chi-Square tests revealed a significant main effect for explicit scaffold condition ($\chi^2(4) = 32.12, p < 0.001$). **Consistent with our design expectations, each explicit scaffold significantly increases the odds of a correct response in the triangular graph task relative to the non-scaffold control.** The (unstandardized) regression coefficients indicate that the two text conditions each roughly double the odds of a correct response, ($e^{\beta_1[\text{conceptual}]} = 2.42, SE = 0.744, p < 0.001; e^{\beta_1[\text{procedural}]} = 2.25, SE = 0.67, p < 0.01$). The static image condition also

¹We chose to model these data at the question rather than participant level because the distribution of total triangular graph scores was bimodal, and an item level model allows us to differentiate between random variance introduced by individual differences in participants and questions, and systematic variance introduced by scaffold condition.

doubles the odds of a correct response, ($e^{\beta_1[\text{staticimage}]} = 2.23, SE = 0.66, p < 0.01$). Most impressively, the interactive image condition increased the odds of a correct response by over a factor of 5, ($e^{\beta_1[\text{interactiveimage}]} = 5.41, SE = 1.61, p < 0.01$).

2.4.2.4 (H4) Graph Order as Scaffold

Wald Chi-Square tests did not indicate a significant main effect for **graph-order** ($\chi^2(1) = 1.92, p = 0.17$). **Contrary to our (H4) hypothesis, completing the TM graph task after the LM graph task did not significantly improve performance with the TM graph.** It is possible that in order to glean salient differences between the TM and LM graphs, they need to be viewed simultaneously (as in Figure 2.8 or Figure 1.11).

2.4.2.5 Effect of Scenario.

As our repeated measures design (participants repeating the graph reading task once with each graph) necessitated the use of two scenarios, for due diligence we tested for effects of scenario in our statistical model. Unexpectedly, Wald Chi-Square tests revealed a significant main effect for scenario-order ($\chi^2(1) = 32.30, p < 0.001$). Specifically, pairing the event-scheduling scenario B with the TM graph task increased the odds of a correct response by a factor of 3, ($e^{\beta_1[\text{scenarioB}]} = 2.97, SE = 0.57, p < 0.001$). When answering questions in the “task scheduling” scenario A ($M = 9.20, SD = 4.12$), participants had significantly lower scores compared to the “events scheduling” scenario B ($M = 10.52, SD = 2.97$). In an online pilot we found no significant differences in performance between the scenarios when tested with the LM graph. To explore the source of this effect, we examined the data sets constructed for each scenario, and in particular, the very first question students solved with the TM graph. In the “task scheduling” scenario A (Figure 2.10, *left*) we see that if a learner makes the most common mistake—seeking an orthogonal intersection from the x-axis—there is a single data point that intersects the line: an available answer. However, in the “events scheduling” scenario B (Figure 2.10, *right*), there is no intersecting data point. We hypothesize that students who were randomly assigned

to this second scenario received implicit feedback that they were misreading the graph if they sought the orthogonal intersect because there was no answer to the question. We suspect this drove students to re-evaluate their strategy, yielding significantly higher scores for the “events scheduling” scenario. We explore this idea further in Chapter 3.

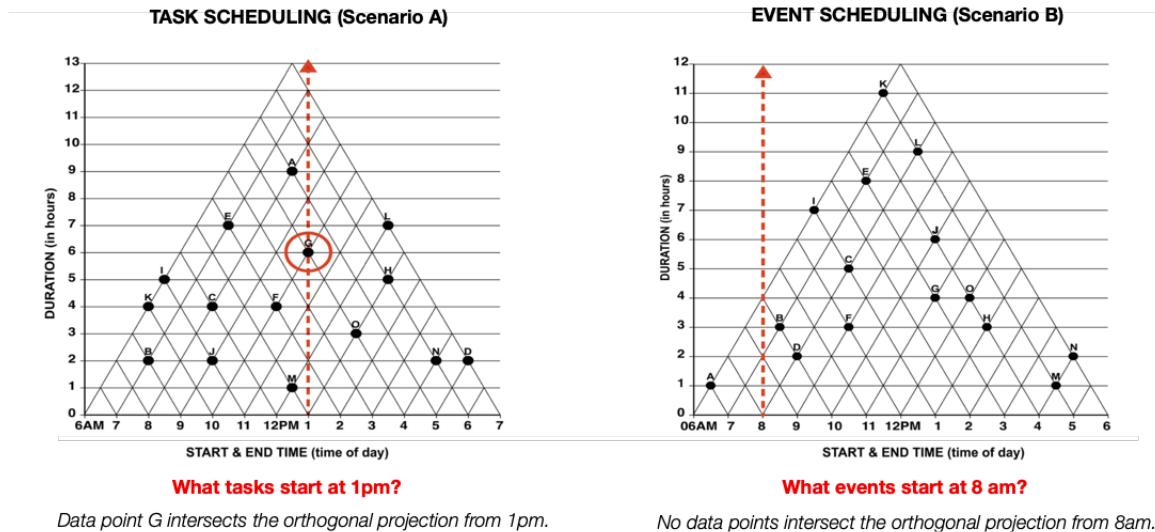


Figure 2.10. Study 2 (Results) — Non-Equivalence of Scenarios. The first question in the Task Scheduling Scenario (*left*) has a datapoint intersecting the orthogonal projection from the x-axis, thus providing a potential ‘orthogonal answer’, but the first question in the Event Scheduling scenario (*right*) does not.

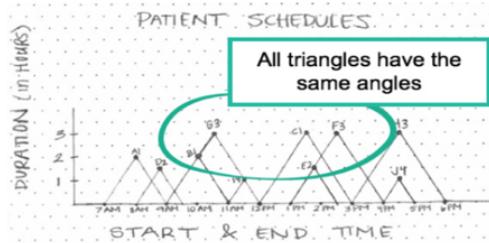
2.4.2.6 Graph Drawing Task

The graph drawing tasks allows us how to explore how each scaffold supports students learning the graphical framework of the TM. We expect that accurately drawing requires deeper understanding of how the graph works, and analysis of any systematic mistakes students make in drawing may reveal sources of difficulty in comprehension. Following the directed approach to qualitative content analysis (Hsieh and Shannon, 2005), a team of 3 raters classified all 316 drawings first into three distinct categories defined *a priori*: [triangular, linear, other], and subsequently into five distinct categories based on the data present in the sample: (correct) triangular, linear, scatterplot, “asymmetric triangular” and “right-angled”. Inter-rater reliability

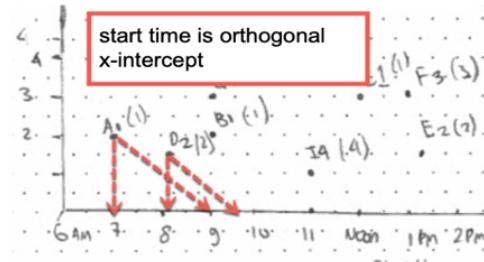
was high ($\alpha = 0.96$) and disagreements were resolved through negotiation, such that the final assignment of categories reflects rater consensus.

Despite relatively low scores on the TM graph reading task, the majority (73%) of participants drew correct TM graphs. 17 individuals (5%) constructed LM graphs, while 3 participants drew scatterplots with start & end time on the x/y axes respectively. Most interesting were the two alternative triangular forms constructed by 66 (21%) individuals: right-angle triangle, and asymmetric triangles (described in Figure 2.11). At top left (a) we see correct triangular graphs, produced by 230 students (73% participants). Alternatively, in (b) *at top right* 44 students drew “right-angle” graphs. They plot duration on the Y axis and the interval as a point, but mistakenly use an orthogonal x-intercept for start time. In (c) *bottom left* 22 students drew forms that were also triangular (plotting the orthogonal intersection as the midpoint of the interval), but the triangles were not geometrically similar because duration was not on the y-axis. Only 17 students drew LM graphs (d) *bottom right*.

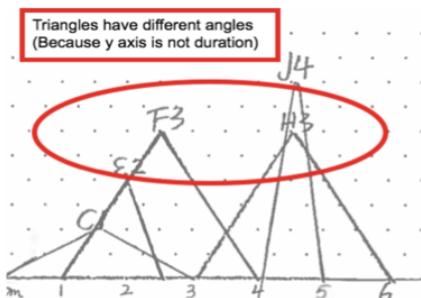
(a) accurate, triangular model graphs



(b) right-angle triangles



(c) asymmetric triangles



(d) linear model graphs

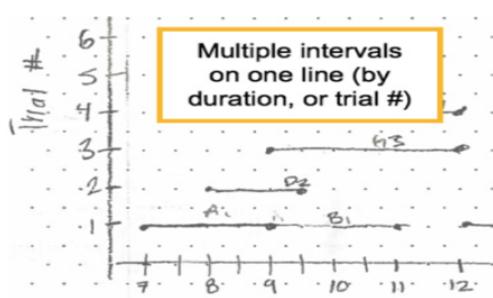


Figure 2.11. Study 2 (Results) Drawing Types. We classified participant drawings into four types based on how the allocated interval properties to each axis.

2.4.3 Study 2 Discussion

The results of Study Two leave us with a conundrum: why were the scaffolds designed by learners in Study One largely ineffective? None of our designs replicated the results of Qiang, Delafontaine, Versichele, De Maeyer, and Van de Weghe, 2012 which yielded better performance with the TM than LM graph, though there were notable differences in our tasks, including their use of an interactive graph interface with hundreds of data points, and feedback in the video instruction. Setting aside the differences in performance between the LM and TM graphs, we assessed the efficacy of scaffold designs by looking at TM scores alone. The widely-held assertion of Study One participants that simple text and image instructions would dramatically improve readability of the graph were not borne out, as on average, participants who received the static scaffolds performed no better than those who received (as participants in Study One) no graph instructions at all.

We suspect the source of this discrepancy lies in a *hindsight bias*. Once students understand how the graph works, they cannot “unsee” it, and therefore underestimate the strength of their prior expectations. The unexpected effect of scenario on TM scores supports this interpretation, as students who received implicit feedback they were reading the graph incorrectly (because there was no available answer) performed better than those who did not (Figure 2.10 right vs. left). In this way, the structure of the task presented the reader with a form of *mental impasse* where their expectations (based on prior knowledge of Cartesian graph forms) left them with no solution, and their attention was actively redirected to reconsidering these expectations. The role of attention can also address why the interactive image was superior to the static text and image scaffolds. If it is the case that a reader does not realize they are misreading the graph (as we observed in Study One), it is easy to ignore the static scaffolds. However, it is much more difficult to ignore a stimulus that appears every time the mouse is moved over a data point.

2.5 General Discussion

While the Triangular Model (TM) graph is elegant in its simplicity, the results of these studies demonstrate this simplicity is deceptive. Without assistance, most readers misinterpret the graph as the conventional representation for temporal intervals: the Linear Model. Even with explicit cognitive aids, many students persist in this erroneous interpretation.

These results have implications for both the design of scaffolds and of new unconventional graphs. First, when designing scaffolds one should consider the reader's expectations based on the conventional representation for variables *in the domain of data being represented*. It is from that prior knowledge that readers begin their interpretation, not from a blank-slate (i.e. a generic graph schema) as we might expect based on a graph's surface features. To overcome this, our results suggest that techniques actively directing attention to salient differences will prove most effective. The interactive-image scaffold achieves this through repeated, user-driven exposure to the multiple intersections of a TM data point with the x-axis. Similarly, we believe the mental impasse provided by the questions in our event-scheduling scenario actively directed readers' attention to their mistaken interpretation. We explore this phenomenon further in Chapter 3

When constructing unconventional graphs, a designer's priority is the computational affordances making the new graph-form suitable to the intersection of data, task, audience, and communicative context. But as we learn from these studies, a designer should also ask, "What expectations will be invoked by the marks on the page?" For the TM graph, we suspect it is the orthogonal axes that drive readers to expect a single orthogonal intersection for each data point. But there is —strictly speaking —no reason that the axes *need* to be orthogonal. In fact, one clever participant in our graph drawing task produced what we believe to be a substantial improvement upon the TM graph, where the y axis was positioned diagonally on the left side of the graph's bounding triangle. We explore this design alternative, and the influences of gridlines, marks and orientation in Chapter 4

In this chapter, we have explored only a small subsection of the design space of scaf-

folding techniques for this particular kind of unconventional graph. We expect our conclusions will generalize to other novel coordinate systems. Our choice of scaffolds was inspired by direct observation and participatory design, however, we suspect a wider range of techniques might be effective in more instructional settings, including explication of worked examples, or seeing the graph being drawn. While we chose to separate our text and image scaffolds to test their differential efficacy, a combined text/image annotation could prove effective even in static media.

We started by reasoning that existing scaffolding techniques would be insufficient for unconventional graphs because learners would lack the prior knowledge of the new graph system required to make use of them. As Pinker (1990) suggests, when confronted with an unfamiliar graph form, the reader instantiates a generic “general graph schema”. But we do not know what the contents of this schema are, or how it guides interpretation. Our results suggest it is possible for the reader to instantiate a schema that actively conflicts with the surface structure of the marks on the page. The novelty of the salient diagonal gridlines in the TM graph was not enough for most learners to suspend their Cartesian expectations. To overcome this prior knowledge, we argue that successful scaffolds for unconventional graphs must not only show or tell us how to read them, but to rather alert us that that we need to pay attention, and reconsider our expectations in the first place.

Acknowledgements

This chapter includes material as it appears in Fox and Hollan, 2018. *Read It This Way: Scaffolding Comprehension for Unconventional Statistical Graphs*. In P. Chapman, G. Stapleton, A. Moktefi, S. Perez-Kriz, & F. Bellucci (Eds.), *Diagrammatic Representation and Inference* (pp. 441–457). Springer International Publishing. The dissertation author was the primary investigator and author of this paper.

Chapter 3

Explorations of Insight Problem Solving

We've learned that despite the efficiency with which the Triangular Model (TM) represents interval relations, it lacks *discoverability*. That is, most readers without prior exposure struggle to interpret its novel coordinate system. Inspired by research in problem solving, in this chapter we ask whether treating an unconventional graph as a type of *insight problem* might yield new effective methods for scaffolding discovery.

In Study 3A, we test whether intentionally imposing a mental impasse improves interpretation, and find that readers faced with our experimental manipulation are more likely to produce (accurate) triangular responses, as well as other types of non-orthogonal responses to TM graph reading questions. In Study 3B, we replicate and extend these findings to compare the effectiveness of the impasse structure to the image-based scaffolds from Study 2. We find that both approaches significantly improve comprehension, and that the interventions have an additive effect. In Study 3C we explore the role of working memory in graphical discovery, finding that the impasse structure is most effective for readers with high working memory capacity. Taken together, these results support our claim that it is fruitful to consider the comprehension of novel graphs through the theoretical lens of problem solving, yielding implications for the design of cognitive aids for novel graphical forms, and the way we characterize the constituent processes of graph comprehension.

3.1 Problem Solving and Insight

In our first observational study with the Triangular Model (TM) of Interval Relations (Study 1), most participants exhibited an *orthogonality bias*: systematically misinterpreting the graph as a cartesian scatterplot by presuming that the relationship between any given data point and the x-axis is defined by single orthogonal projection between the two. But we don't believe this is a failure in *perception*: that readers did not *see* the diagonal gridlines. Rather, we argue it is more effective to construe this as a failure of *interpretation*: readers did not know what meaning to assign to the gridlines. They were unable to "solve the problem" of how the marks were related in space.

For a few successful outliers however, their production of the correct interpretation was accompanied by a protracted struggle, a sudden clap of their hands and ecstatic exclamation, "***Oh! That's how it works!***" What we observed were moments of **insight**.

In a 1992 contribution reconciling contemporary research on problem solving with information-processing accounts of cognition, Stellan Ohlsson offered a new conceptualization of the term 'insight'. Ohlsson introduced the concept of **impasse**, suggesting that rather than a sudden appearance in consciousness of a complete, correct solution to a problem, insight is better operationalized as what occurs after a problem solver breaks free from an impasse: "*a mental state in which problem-solving has come to a halt; all possibilities have been exhausted and the problem-solver cannot think of any way to proceed*" (Ohlsson, 1992, pg. 4). The implications of this insight on insight were substantial, affording an information-processing account of the phenomenon consistent with empirical research that found: (1) individual instances of insight did not necessarily guarantee a correct solution to a problem, and (2) a fully-formed solution did not always appear directly following a moment of insight. Rather, a problem solver might require multiple, successive insights to find a correct solution, and for some, a correct solution might never appear.

But what leads a problem solver to experience an impasse in the first place? One

related construct (predating Ohlsson's conceptualization) that offers an answer in the context of some problems is *functional fixedness*: the idea that the experience of using an object in a particular way lowers the probability of finding a solution in which the object is used in a *different* way. Examples of functional fixedness abound in classic insight problems. In Dunker's candle problem (1945), the solver is fixed on the function of the material BOX as CONTAINER, ostensibly lowering the probability they envision the (correct) solution using BOX as PLATFORM. Similarly, in the two-string problem, solvers are fixated on the function of PLIERS as TOOL rather than the desired solution of PLIERS as PENDULUM BOB. When a problem solver is fixated on the conventional function of an object and unable to envision an alternative use, they attempt suboptimal (and ultimately incorrect) solutions, leading to a state of impasse. According to Ohlsson (1992) it is only after reaching this impasse that the solver is likely to restructure their (mental) representation of the problem, allowing them to derive a correct solution.

In the case of graph discovery, it is reasonable to expect that substantial experience using the most common forms of visualizations (scatterplots, bar charts, and line graphs) may serve to fix our expectations of axes and their underlying coordinate system toward a cartesian interpretation, where a point is defined uniquely by a pair of numerical coordinates derived via its orthogonal intersections with a horizontal and vertical axis. We argue that the orthogonality bias exhibited by many TM graph readers can be reconstrued as a sort of *graphical fixedness*, where readers are fixated on the relationship between the point and x-axis as orthogonal—unable to conceive of an alternative relationship between these marks in space.

3.1.1 Constructing a Mental Impasse

Unlike many classic problems in the insight literature however, we cannot be certain that readers of even an unconventional graph will ever reach a state of impasse. In Study 1 we found that most unsuccessful graph readers were unaware they had failed to solve the problems correctly. That is, there was an *available* (though incorrect) solution to the graph-reading problem, and

thus the reader did not necessarily reach a state where they were stuck, unable to think of a way to proceed. This would be akin to one of the suboptimal attempts at the candle problem (such as trying to use a pin or tack to attach the candle to the wall) being considered by the problem solver as an acceptable solution.

Thus, in order to treat the interpretation of an unconventional coordinate system as an insight problem, we must *intentionally craft* a state of impasse in order to draw a reader's attention to their most probable misconception. In this work we will refer to this **as imposing an impasse-structure**—constructing a situation which increases the probability of the solver experiencing a mental impasse.

The most obvious way to invoke a state of impasse for the TM graph is to offer a reader feedback that their solution is incorrect. In Study 1, we gave verbal feedback during the debrief interview prior to the scaffold design task (e.g. “Your answer to question 1 was incorrect. Would you like to take another look?”) In many cases, this feedback was sufficient for the learner to reconsider, correct their interpretation, and arrive at a correct solution. Simple feedback appears to be very effective, and was also utilized by Qiang & colleagues (2012) in their evaluation study of the TM graph, where they used a combination of video instruction and interactive feedback to train students how to read the graph, and eventually utilize its computational efficiency to answer questions more quickly and accurately than with the informationally-equivalent but conventional alternative, the Linear Model. But this sort of feedback is not possible in a self-directed learning situation, such as a reader of scholarly paper who encounters an unfamiliar graph communicating study results. In Study 2, however, we inadvertently found a method for creating an impasse-structure without didactic feedback, when after finding an unexpected effect of task scenario (which determined the configuration of data points visualized in the TM graph) a post-hoc analysis revealed that one of the scenarios did not have an available answer to the first question *if* the reader was misreading the graph as a cartesian scatterplot (see section 2.4.2.5). This coincidence in the design of our materials meant that for readers randomly assigned to that scenario, who tried to read the TM graph as a scatterplot, there was no datapoint intersecting an

orthogonal projection from the x-axis. We believe this presented readers with an obstacle: ‘there is no answer to this question’ thus invoking a state of *impasse*.

3.2 Research Goals

In Studies 3A-C we systematically explore the potential role of *mental impasse* in discovering the coordinate system of a novel graph.

In Study 3A we test the impasse hypothesis, and determine if our findings in Study 2 (an unexpected effect of scenario) are in fact attributable to invoking an impasse state, or rather some other attribute of the stimulus scenarios. We replicate these findings in an online study, and then in Study 3B directly compare the effectiveness of explicit scaffolding (static and interactive worked-example images) verses the sort of implicit scaffolding we argue is offered by imposing an impasse-structure. In Study 3C we explore a potential source of the substantial individual differences in graph interpretation that exist (even with the support of implicit and explicit scaffolding) by evaluating the relationship between task performance and working memory capacity.

3.3 Study 3A: Testing the Mental Impasse

We begin our evaluation of the impasse hypothesis by directly comparing performance on an interval graph reading task with the Triangular Model (TM) graph for participants who receive materials that are *intentionally designed* to invoke a mental impasse, against those who receive materials intentionally designed to *not* pose a mental impasse. We argue that we can lead readers to state of impasse by carefully designing the question paired with a particular underlying dataset (i.e. the position of points) represented in the TM graph such that there is **no data point intersecting an orthogonal projection from the x-axis warranted by a cartesian interpretation of the coordinate system**.

Procedurally, this means that for any given problem, if a reader attempts to extract the

interval relation posed in the question by interpreting the coordinate system as cartesian, they will **not** find a datapoint intersecting their (visual saccade, finger or mouse-cursor tracing) of the graph: they will find there is no available answer to the question (*see Figure 3.1 right*). Alternatively, readers in a (non-impasse) control condition posed with the same problem **will** find a data point intersecting the corresponding orthogonal projection (see Figure 3.1 left). We argue that finding an available answer to the question means that readers are likely to select that

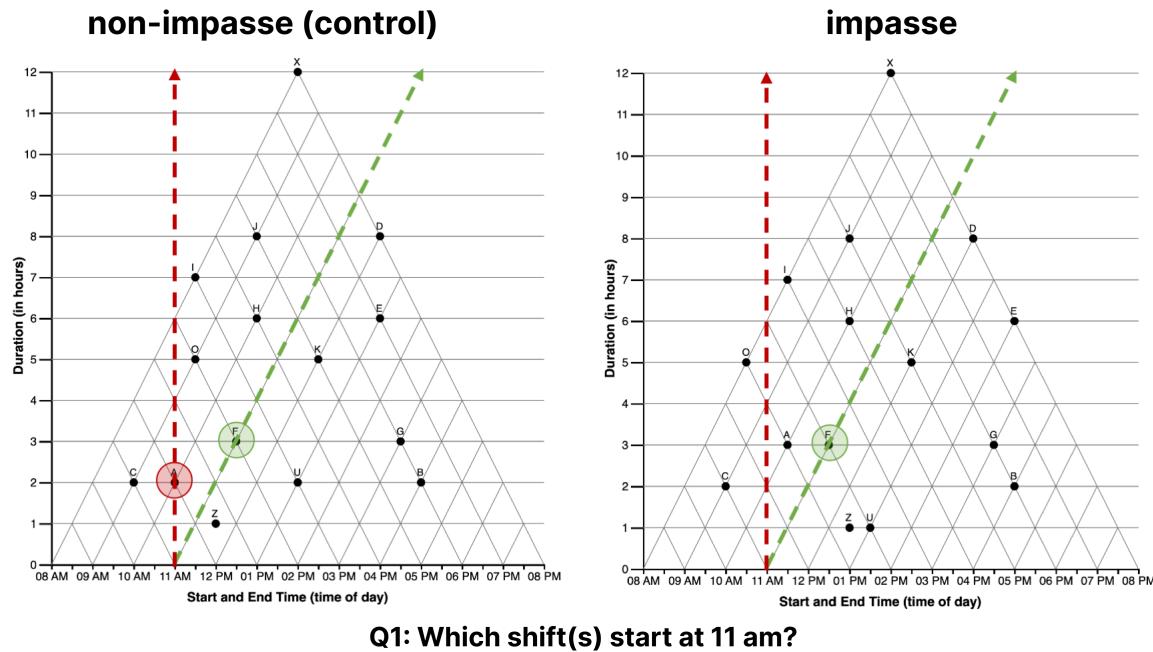


Figure 3.1. Study 3A Experimental Conditions. For the question “*Which shift(s) start at 11 AM?*” the stimulus graph for the control condition is displayed on the left, and the impasse condition on the right (with red and green lines added for explanation; these do not appear in the actual stimuli). Note that across both conditions, the question and design of the graph (position and orientation of axes, gridlines and labels) are identical. The conditions differ *only* in the position of the datapoints that are represented. In both conditions, the green dotted line indicates the optimal path to the correct answer: starting at the reference time [11AM] on the x-axis, and tracing up the *ascending* diagonal gridline (in green; which indicates start time). Only point [F] intersects this line, and is the correct (triangular) response. Alternatively, if the graph is read as a cartesian scatterplot, the reader would follow a suboptimal path to an incorrect answer: starting at the reference time [11AM] on the x-axis, and **projecting an non-existent orthogonal line** (in red) through the graph space. Note that in the control condition, the datapoint [A] intersects this line, thus providing the reader with a potential response to the question. In the impasse condition, however, there is **no datapoint** intersecting this line, and therefore no available answer to the question.

answer as their response, while alternatively *not* finding an available answer to the question will present the reader with an obstacle. The reader has a reasonable expectation that there should be an answer to the problem, and when they cannot find one, they will enter a state of *impasse*.

Specifically, we hypothesize that:

- (H1) Participants in the *impasse* condition will have have a higher probability of a correct response than participants in the *non-impasse* (control) condition.
- (H2) Participants in the *impasse* condition will provide more transitional (i.e. incorrect, but non-orthogonal) responses than participants in the *non-impasse* (control) condition.

3.3.1 Methods

3.3.1.1 Participants

We recruited 146 undergraduate students at UC San Diego to participate (in person) in exchange for course credit. Twenty participants were excluded for failing attention-check questions, yielding 126 participants for analysis (gender: 37 % male, 62 % female, 1 % other; age: 18 - 33 years).

3.3.1.2 Design

The experiment employed a multilevel design structure with 1 fixed, and 2 random factors:

- (F1) **implicit scaffold** (between-subjects) @ (c = 2) levels : *none* [control], *impasse*
- (R1) **question** (within-subjects) @ (q = 13) levels
- (R2) **participant** @ (n = 126) levels

Participants were nested within implicit scaffold condition, and questions were fully crossed with condition. Thus, each participant was randomly assigned to one implicit scaffold, in which they completed all the questions (with the TM graph).

3.3.1.3 Materials

Interval Graph Comprehension Task

The Interval Graph Comprehension Task is an extension of the tasks used in Studies 1 and 2, with two improvements. First, we simplified the interval relations in the questions to improve construct validity and ensure that performance on the task primarily reflects interpretation of the coordinate system, and not a participant's ability to perform algebra over interval relations (see Allen's Interval Logic). Secondly, we carefully designed the combination of question/response options/visualized dataset to be **interpretation discriminative**. An answer designated as correct can never be produced by both an orthogonal and a triangular interpretation of the coordinate system¹. This has the effect of pushing total task accuracy scores toward floor (for orthogonal) and ceiling (for triangular) interpretations.

The task begins by situating participants in a problem solving scenario where they are to assume the role of factory manager responsible for scheduling employee work shifts. They are instructed to complete the task by using a graph of the schedule to answer questions about the timing of shifts. A shift is defined as an interval of time with a discrete start and end time (on the hour). Thus the datapoints on the TM graph correspond to shifts, and shifts are identified with letters (i.e. A, B, C, etc). We chose to situate the task in the context of a scenario to give participants a familiar conceptual anchor for the interpretation of datapoints. Rather than the abstraction of 'an interval of time' each datapoint refers to a shift in an employee's schedule.

The task proceeds with two experimental blocks, consisting of 5 and 10 ordered items respectively. Within each block, participants complete the items in sequence, without feedback, and do not have the ability to skip ahead, nor return to previous items. In each item, the participant is presented with a TM graph, a question, and a grid of response options (see Figure 3.2). The design of the TM graph (shape, scale, axes, labels, size and relative screen position) are identical across items and conditions. The stimulus web application renders the graph as 700x700 pixels.

¹13 of the 15 questions have this discriminative property. Questions #6 and #9 do not, and were used for attention check and exploratory analyses respectively. They were not included in analyses for hypothesis testing.

It detects browser window size and forces the browser into full screen mode. If the screen size is below the minimal threshold, participants are prevented from starting the task.

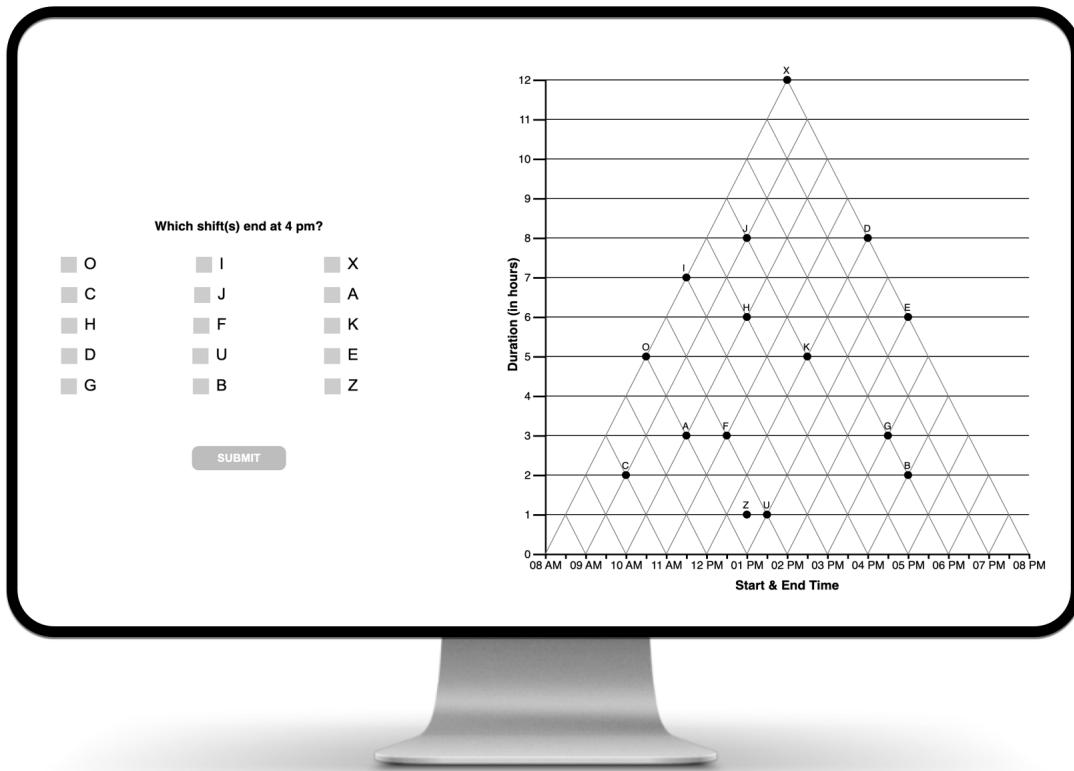


Figure 3.2. Study 3A – Layout of Graph Interpretation Task. See above the layout of the Interval Graph Interpretation Task (shown is Question #4; impasse condition). Each item consists of a TM graph (at right), accompanied by a question, and grid of response options. Participants were instructed to check the boxes corresponding to all of the point in the graph that answered the question, and click the SUBMIT button to proceed to the next question.

Across both experimental conditions, participants complete the same sequence of blocks with identical questions. The experimental manipulation of **implicit scaffold** is accomplished by rendering a different underlying set of data (i.e. a different shift schedule) for each condition. The shift schedule visualized for the *impasse* condition imposes an impasse-structure as described in Section 3.3: for each question, there is no datapoint intersecting the orthogonal projection from the x-axis warranted by a cartesian interpretation of the coordinate system. Figure 3.1 displays the TM graph (and thus corresponding shift schedule) displayed for the first experimental block

in each condition (*control* on left, impasse on right; green and red lines are not present in the stimuli). Note that the position of corresponding data points (e.g. shifts labelled as A) across the two conditions are very similar. We carefully engineered the *impasse* condition by strategically moving specific data points a minimal distance away from their position in the control condition graph. We did this to minimize the potential for differences in accuracy or response time that might be introduced by dramatically different optimal scan/tracing-paths as an effect of datapoint position. For example, if the correct response to Q1 in the control condition requires the reader to trace halfway up the diagonal gridline, but the impasse condition requires them to trace to the top of the graph (perhaps crossing additional datapoints) then performance differences may not be solely attributable to the impasse manipulation. This minimal difference in datapoint position also allows us to more directly compare any mouse-cursor movement the participant may make while answering the questions.

Each question in the task corresponds to one or more interval properties or relations, with questions increasing in complexity as the task progresses. The first question involves the *start-time* property: Which shift(s) start at 11 am? To answer the question, the participant must locate all points intersecting the 11am gridline of the graph, and then check the corresponding check-boxes on the left side of the screen. In the first experimental block, the easiest questions were posed, only requiring extraction of interval properties (i.e. start-time, end-time). At the end of the first block, participants see a second task instruction screen, progressing the scenario by informing them they will now answer questions about the schedule for the following week (an interruption giving for changing the schedule of shifts being visualized for the second experimental block). In this second block, the two experimental conditions converge, utilizing identical graphs, datasets and questions (i.e. the impasse manipulation is only applied to the first experimental block.) Questions in the second block increase in difficulty, asking about relations between intervals. For example (Q8): Which shifts less than 7 hours long start before B begins and end after X ends? This requires identification of *start-time*, *end-time*, and *duration*, as well as the *before* and *after* relations. After completing the second block of items, the scenario is

concluded. The entire time on task ranged from 6 to 24 minutes.

Measures and Scoring The type of item employed in the Interval Graph Comprehension Task (as seen in Figure 3.2) is described in the educational tests and measures literature as Multiple-Response (MR), or Multiple Choice Multiple Answer (MCMA). In a traditional (Single Answer) Multiple Choice (SAMC) question, a respondent marks a single response from a limited option set. One point is given for selecting the option designated as correct, and zero points given for marking any of the alternative (i.e. distractor) options. In MCMA questions, however, the set of response options selected by the participant (termed their *answer*) might be *partially* correct. They may have correctly selected some options, incorrectly selected other options, and incorrectly not selected other options. In this way, MCMA item-types offer more information to the researcher than traditional multiple choice items, offering insight into the nature of errors made by the respondent. The tradeoff, however is that for MCMA items, it is not obvious how to allocate partial credit. A number of scoring schemes have been evaluated in the literature, and systematic comparison is offered by Schmidt & colleagues (2021). They conclude that it is the task of the researcher to select the scoring scheme that offers the greatest discriminative ability based on the goals of the underlying test.

The MCMA item type is well suited for the goals of this task as it gives us insight into how the respondent interprets the coordinate system (via their selection of specific data points) without nudging the reader toward particular solutions (as would be the case with a limited option set in a traditional multiple choice measure). To meet our analysis and hypothesis testing goals, we utilize two scoring approaches described by (Schmidt, Raupach, Wiegand, Herrmann, and Kanzow, 2021) and derive two measures to characterize a participant's response on each question.

ACCURACY is a binomial dependent variable (*0: incorrect, 1: correct,*) indicating if a given response is correct according to a triangular interpretation of the coordinate system (using the *absolute scoring* scheme [Schmidt, Raupach, Wiegand, Herrmann, and Kanzow, 2021

Method #1]). It is the most conservative measure of performance based on the item type, in that there is only one unique combination of the ($n = 15$) response options that yields a *correct* answer, out of 2^{15} possible response combinations. There are many more ways to get an accuracy score of 0, than to get an accuracy score of 1. The ACCURACY measure tells us if a participant has arrived at a correct interpretation of the coordinate system for that particular question. But if they have not (i.e. a score of 0) it does not tell us anything else about the nature of their understanding.

To address this question, we derive a second dependent variable we call INTERPRETATION, indicating which alternative interpretation or interpretative strategy the response most closely matches. The options of this 4-level (ordered) factor are: *orthogonal*, *other*, *angular*, and *triangular*. Responses labelled as *orthogonal*, *other* and *angular* are technically incorrect (i.e. receive ACCURACY scores of 0), but they are incorrect in distinct and empirically interesting ways.

- *triangular* : includes only correct triangular responses. These responses indicate a correct interpretation of the triangular coordinates in the context of the given question
- *angular* : includes responses that isolate data points along diagonal or horizontal gridlines connecting the reference point in the question, as well as cases where the participant selects *both* orthogonal and triangular answers. These responses indicates some degree of angular/triangular coordinate understanding, or uncertainty, but are not strictly correct
- *other* : includes blank responses, cases when the reference point is selected (i.e. datapoint referenced in the question, such as, "What shifts start at the same time as D"; where D is the reference point) , and other responses that can't be classified (including selecting all datapoints). These responses indicate an uncertain or unidentifiable interpretation, but one that is distinctly not orthogonal nor triangular
- *orthogonal* : includes the orthogonal-consistent responses designated for each question.

We derived these responses by superimposing a linear model graph atop the TM model (reading the TM as a cartesian scatterplot). Also included in this category are *satisficing* attempts at orthogonal answers, produced by selecting the nearest points to the orthogonal projection (in the *impasse* condition, when an orthogonal intersecting data point is not available). These responses indicate a primarily orthogonal/cartesian interpretation.

Figure 3.3 displays examples of each INTERPRETATION for the second question based on actual responses in Study 3A. For a full description of the how the INTERPRETATION measure is derived using a partial $[-1/q, 0, +1/p]$ scoring scheme in combination with interpretation-specific answer keys, refer to Appendix D.

3.3.1.4 Procedure

Participants completed the study in person, seated at a desktop computer where they interacted with a custom web-application² via the Chrome web-browser, keyboard and external mouse. After agreeing to an IRB-approved informed consent, participants were randomly assigned to an **implicit scaffold** condition and presented with task instructions. They then completed the Interval Graph Comprehension Task. Upon completion, participants were presented with a series of questions about their effort and enjoyment of the task, followed by a demographic questionnaire, and final debriefing text.

3.3.1.5 Analysis

Response Accuracy

To test hypotheses related to response ACCURACY we fit mixed logistic regression models (generalized linear mixed models (GLMM) with a logistic link function) in R using the `lme4` package (Bates, Mächler, Bolker, and Walker, 2015; R Core Team, 2022). Note that we choose to model these data at the question rather than participant (i.e. TOTAL SCORE) level because the structure of a mixed effects model allows us to differentiate between random variance introduced

²implemented with the JsPsych framework de Leeuw, 2014

Question 2 : Which event(s) start with D?

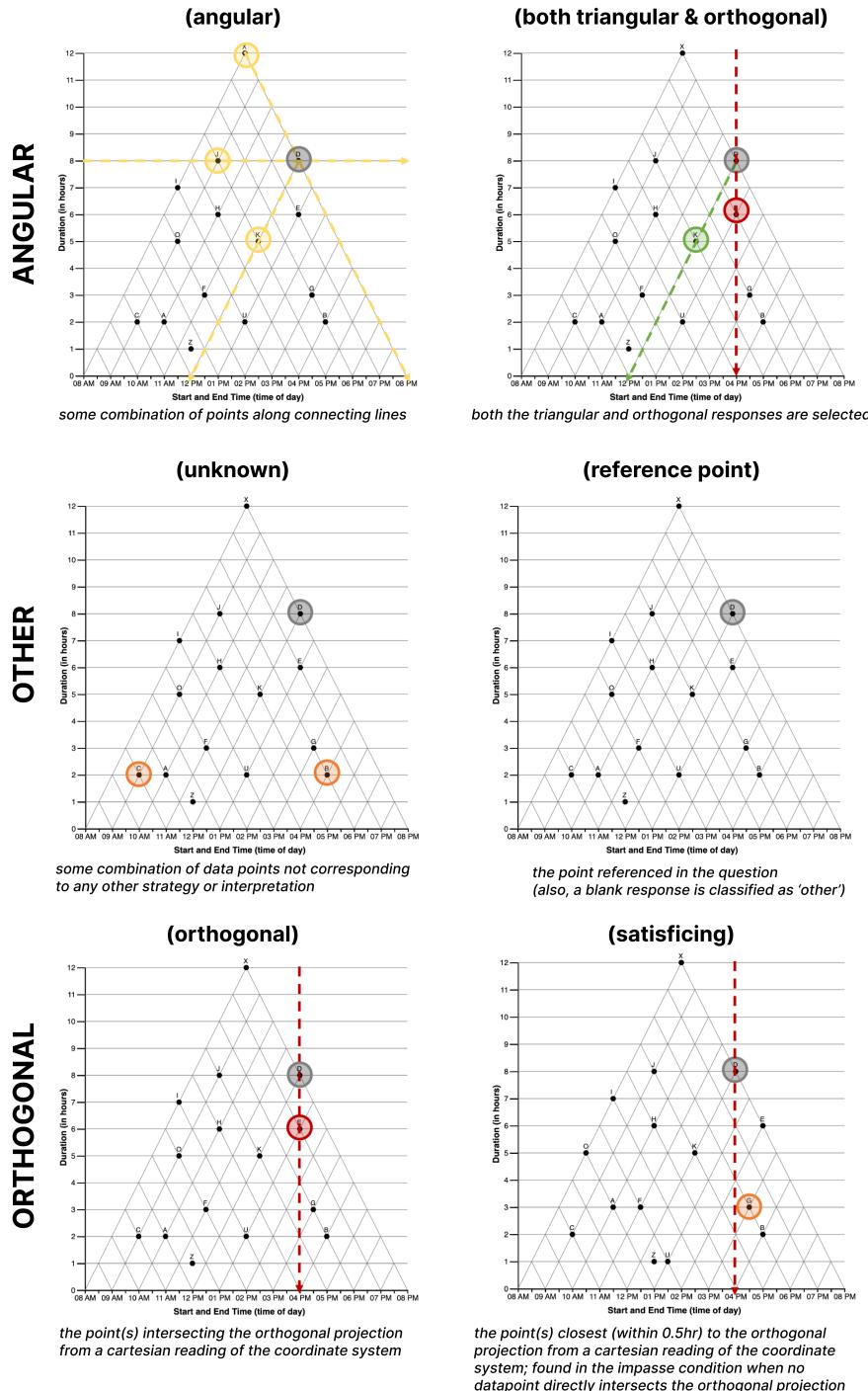


Figure 3.3. Study 3A – Examples of Interpretation Measure. For Q2 of the task (Which events start with D?), the (correct) *triangular* interpretation is given by data point K. Additional alternative interpretations are described, above.

by individual participants and questions, versus the expected systematic variance of experimental condition. Further, the distribution of total accuracy score at the participant level was bimodal, and violated the assumptions of normally distributed residuals and homogeneity of variance required by OLS linear regression. For contrast coding categorical variables, the default treatment (dummy) coding scheme was used: on the response variable ACCURACY the level *0:incorrect* was defined as the reference category, and on the predictor variable **implicit scaffold** the level *non-impasse* (control) was defined as the reference category. Thus exponentiated model intercept e^{b_0} refers to the baseline odds of a *correct* response in the *non-impasse* (control) condition, while exponentiated model coefficient e^{b_1} refers to the odds-ratio (relative increase or decrease in odds) of a *correct* response in the *impasse* condition relative to the *non-impasse* (control). To determine a final model we first defined the maximal random effects structure theoretically justified by the study design (random intercepts for questions and subjects). We then fit a model with **implicit scaffold** as predictor and used a likelihood ratio test to decide if adding the predictor resulted in a significantly better fit. Statistical significance of each predictor in the final model was determined via Wald Chi-Square tests, and all reported p-values are for non-directional tests with a decision threshold $\alpha = 0.05$.

Response Interpretation

To test hypotheses related to response INTERPRETATION we fit Bayesian³ mixed multinomial regression models in R using the `brms` package (Bürkner, 2017). All models were run with four MCMC sampling chains, a total of 2500 iterations with 1000 warm up iterations. The default treatment-coding scheme was used: on the response variable INTERPRETATION the level *orthogonal* was defined as the reference category, and on the predictor variable **implicit scaffold** the level *non-impasse* (control) was defined as the reference category. The multinomial regression model estimates $(k - 1)$ equations for (k) levels of the response variable. Thus, there

³Although it is possible to fit mixed multinomial regression models under the frequentist framework with the R package `mlogit` we elected to use a Bayesian framework to take advantage of the well-documented `brms` package and its active support community.

is one equation (and therefore set of model estimates) for the relative odds of an *other* (vs *orthogonal*) response, one equation for an *angular* (vs *orthogonal*) response, and one equation for a *triangular* vs. *orthogonal* response. Rather than returning a point estimate for intercepts and model coefficients, the Bayesian model estimates posterior distributions. To facilitate comparison with results for response ACCURACY, we report the median of the posterior distribution for each model's intercept and predictor coefficient, as well as the 95% credible interval, and % probability of direction (*pd*). The *pd* (also known as the Maximum Probability of Effect), varies between 50-100% and can be interpreted as the probability that a given parameter (described by its posterior distribution) is strictly positive or negative. In the context of these analyses, the *pd* value is an indication of whether a given predictor *increases* or *decreases* the odds of a particular response interpretation⁴. To determine a final model we first defined the maximal random effects structure theoretically justified by the study design (random intercepts for questions and subjects). We then fit a model with **implicit scaffold** as predictor and used a Bayes Factor model comparison to determine if there was sufficient evidence in support of the predictor model over the random effects only model. We characterize Bayes Factors using guidelines defined in (Jeffreys, 1961). We also set informative priors for each model. For intercepts, we used a cynical but wide prior on odds of correct response informed by Studies 1 and 2: $\text{normal}(\mu = -1.1, SD = 1.5)$. For the **implicit scaffold** predictor, we used an neutral prior with respect to the direction of the effect (odds increasing or decreasing), with a wide distribution: $\text{normal}(\mu = 0, SD = 2.42)$.

3.3.2 Results

3.3.2.1 Overall Accuracy

To explore the effect of **implicit scaffold** on TM graph reading performance, we start by describing the distribution of TOTAL SCORE, operationalized as the percentage of correct responses (derived from the ACCURACY measure) of the 13 interpretation-discriminant questions

⁴The PD value is similar (i.e. is strongly correlated) to a frequentist p-value, (Makowski, Ben-Shachar, Chen, and Lüdecke, 2019)

of interval graph comprehension task. Across both conditions, TOTAL SCORE ranged from 0 to 100 with a mean of 32%. This low mean is consistent with our findings from Study 1 and Study 2: the TM coordinate system is challenging for learners to independently discover.

STUDY 3A | Distribution of Total Score

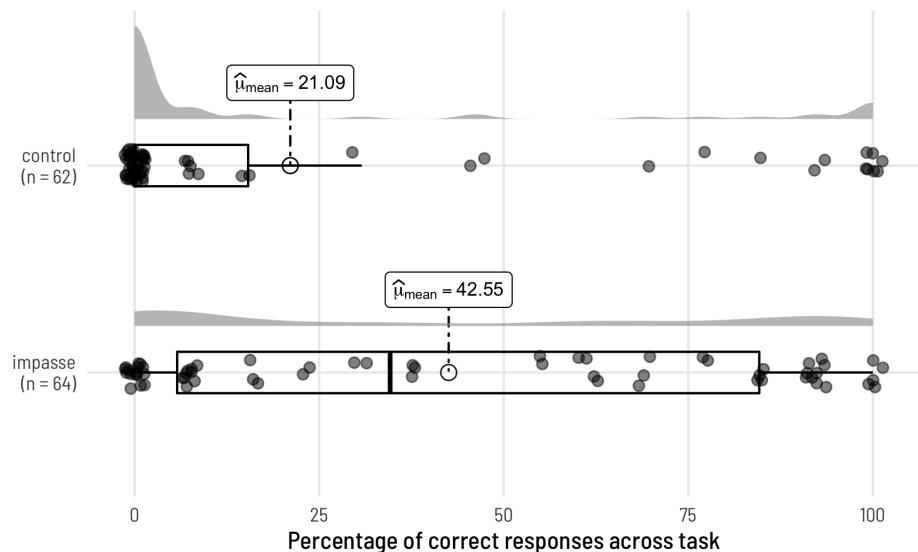


Figure 3.4. Study 3A — Distribution of Total Score. The mean TOTAL SCORE of participants in the *impasse* condition (bottom) is double that of participants in the *non-impasse* control condition (top). There is much greater variance in the *impasse* condition.

In Figure 3.4 we see that the distribution of this outcome variable is bimodal: with modes near the floor (0% correct) and ceiling (100 % correct) of the scale. This bimodality is sensical considering the nature of the task, where each item indexes a different information extraction operation over the same coordinate system. Though some operations may be more complex (involving more steps) or prone to error (requiring selection of more data points), the ability to offer a correct response relies equally on a correct interpretation of the coordinate system. For each item, it is not possible to produce a correct response with a non-triangular interpretation of the relationship between an individual data point, and the x-axis. A score of 100% indicates that the participant correctly interpreted the coordinate system throughout the task, starting at the first question. A score of 0% indicates the individual never correctly interpreted the coordinate

system. A score somewhere in-between indicates that an individual deciphered the coordinate system sometime over the course the task, or that they held a correct interpretation early on but made mistakes (potentially due to carelessness, mistakes in interval logic, or reversion to an incorrect interpretation). One way to conceptualize the effect of **implicit scaffold** on task performance is that it shifts some of the mass of the TOTAL SCORE distribution from floor toward ceiling. That is, posing a mental impasse appears to move some participants away from a score of 0% correct. The effect of the *impasse* condition appears to be individually selective: rather than helping most participants a little, it helps a few participants a lot.

3.3.2.2 Accuracy: Does posing an impasse-structure improve response accuracy?

We test the primary hypothesis that posing a mental impasse will improve discoverability of the TM graph by quantifying the effect of **implicit scaffold** on ACCURACY; operationalized as the probability of a correct response on a given question. Across questions, participants in the *impasse* condition respond reliably more accurately (43% correct) than those in the control condition (only 21% correct), see Figure 3.5 [A]. To evaluate the effect of **implicit scaffold**, we fit a mixed effects logistic regression model with random intercepts for subjects and questions. A likelihood ratio test indicates that a model including a fixed effect of **implicit scaffold** explains significantly more variance in ACCURACY than an intercepts-only baseline model ($\chi^2(3,4) = 17.85, p < 0.001$). The explanatory power of the entire model is substantial (*conditional R*² = 0.89) and the part related to the fixed effect (*marginal R*²) explains 15% of variance.

Consistent with our (H1) hypothesis, the impasse scaffold substantially increases the odds of a correct response. The model estimates that participants in the impasse condition were 62 times more likely to offer a correct response than participants in the control condition ($e^{\beta_1} = 61.9, p < 0.001, 95\% CI [7.21, 531.75]$). Based on the fixed effect of **implicit scaffold**, the model predicts the probability of a correct response in the control condition is effectively 0% (95% CI [8.7e-4, 0.03]), while the probability of a correct response in the impasse condition

STUDY 3A | ACCURACY

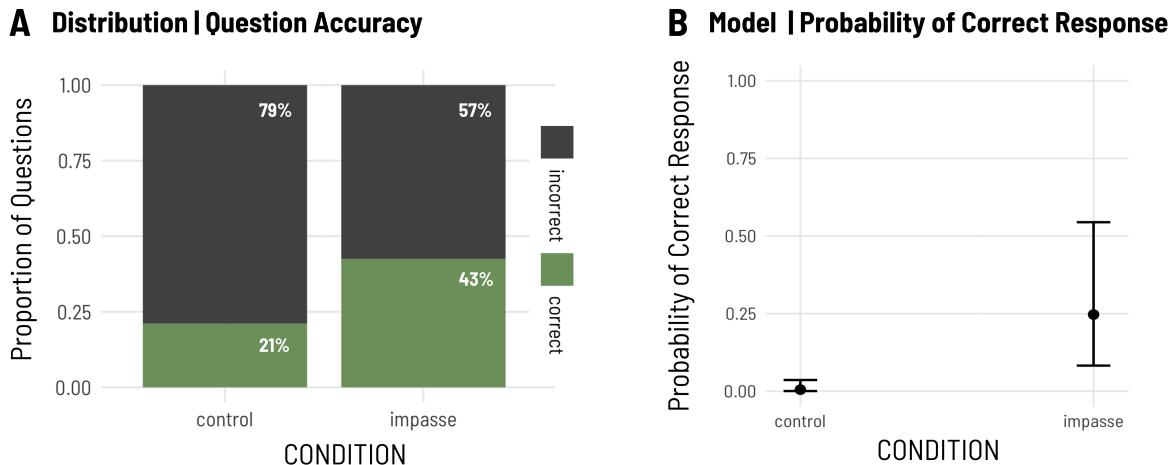


Figure 3.5. Study 3A – Accuracy. [A] A proportional bar chart shows the relative percentage of incorrect (grey) vs. correct (green) responses across all items on the Interval Graph Comprehension Task. [B] The model predicts a significantly higher probability of correct response for the *impasse* condition.

increases to 25% (95% CI [0.07, 0.60]). Raw data are visualized in Figure 3.5 [A], model predictions in Figure 3.5 [B] and parameter estimates and model specification are detailed in Appendix B.1.1.

3.3.2.3 Mouse-Cursor Behaviour

While the ACCURACY score can indicate whether readers correctly interpret the graph, it cannot reveal the strategies employed to answer a particular question. To explore the mechanisms behind our results, we captured mouse tracing data. Similar to eye tracking data, mouse tracing provides an imperfect proxy for visual attention of the learner during the problem-solving session. This is a particularly rich source of insight for our graph reading problems as learners frequently used the mouse to navigate across the graph, the mouse acting like fingers tracing down or across gridlines. Of course, not all learners utilize the mouse to the same extent, and so we limit the present analysis to qualitative observation of gestalt patterns of graph traversal.

Figure 3.6 contains a set of heatmaps generated from raw path and dwell time data depicting the mouse movements of all participants on the first question of the Interval Graph

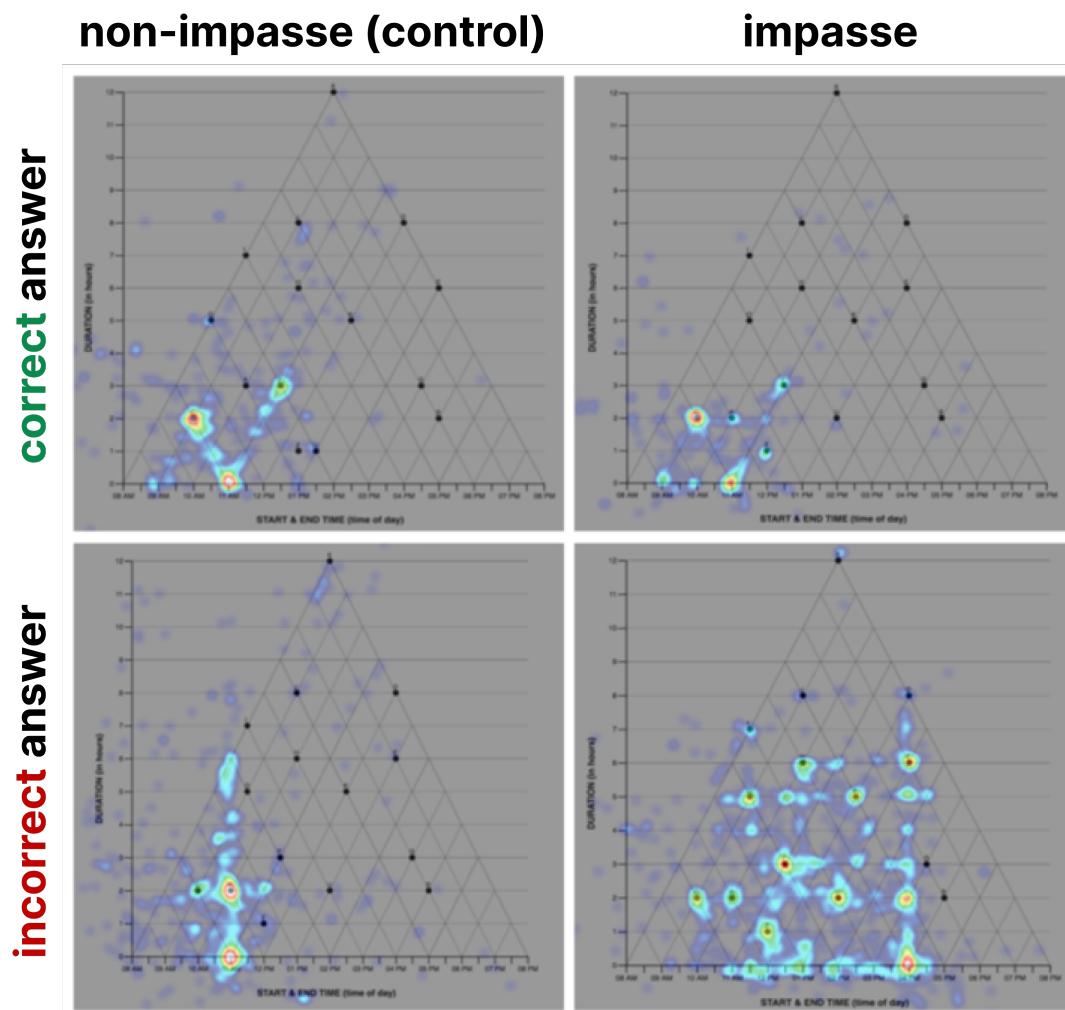


Figure 3.6. Study 3A – Mouse Cursor Behaviour on First Question. Across conditions (columns) participants who offer correct answers on the first question show a consistent (diagonal) pattern of traversal with their mouse cursors. Conversely, each condition yielded very different patterns of traversal for participants who gave incorrect answers.

Reading Task. In the left column, we see data for readers in the *non-impasse* control condition, and on the right, the *impasse* condition. The top row of heatmaps were generated from only those participants who *correctly* answered the question, while the bottom row from participants with *incorrect* answers. Visual inspection of these heatmaps reveal that across both conditions (top row), learners who correctly interpreted the coordinate system traversed the graph in a similar fashion, with the most prominent patterns following the relevant diagonal gridlines, supporting our assumption that these readers develop a correct, *triangular* interpretation of the coordinate

system. Inspecting those with incorrect answers (bottom row), we see dramatically different patterns of tracing across conditions. While those in the control condition (bottom left) follow the expected Cartesian projection, learners in the impasse condition (bottom right) exhibit no single discernible pattern. While these learners did not arrive at the correct answer, the diversity of their tracing behaviour may be an indication of puzzlement, and a variety of intermediate or indeterminate interpretations.

3.3.2.4 Interpretation: Does posing an impasse-structure change graph interpretation?

Although we observe that posing an impasse-structure question substantially increases the odds of a correct response, the overall probability of a correct response is still quite low. Far more participants answer incorrectly than not. As discussed in Section 3.3.1.3 the ACCURACY score does not allow us to differentiate the various ways in which a particular response may be incorrect. For this we need to explore INTERPRETATION: a derived measure indicating which interpretation of the coordinate system a response most closely matches: *orthogonal*, *other*, *angular*, and *triangular*. Differentiating between different kinds of incorrect (i.e. *orthogonal*, *other*, *angular*) responses give us insight into how the impasse effect might work: is the effect of impasse *transitional* (i.e. it helps you restructure out of the orthogonal interpretation), or *absolute* (i.e. it helps you correctly restructure your understanding, or not at all)?

To test the hypothesis that posing a mental impasse doesn't yield only correct triangular interpretations, but also transitional interpretations of the interval-coordinate system, we quantify the effect of **implicit scaffold** on INTERPRETATION by fitting a Bayesian mixed multinomial regression model with random intercepts for subjects and questions. A Bayes Factor model comparison (against a random intercepts-only model) indicates strong evidence for a main effect of **implicit scaffold** ($BF = 1.38e+14$). In Figure 3.7 [A] we see that the *impasse* condition yields not only a lower proportion of *orthogonal* responses, but also a greater proportion of *other*, a small increase in *angular*, and substantial increase in correct *triangular* responses.

Consistent with our (H2) hypothesis, the impasse condition substantially increases

STUDY 3A | INTERPRETATION

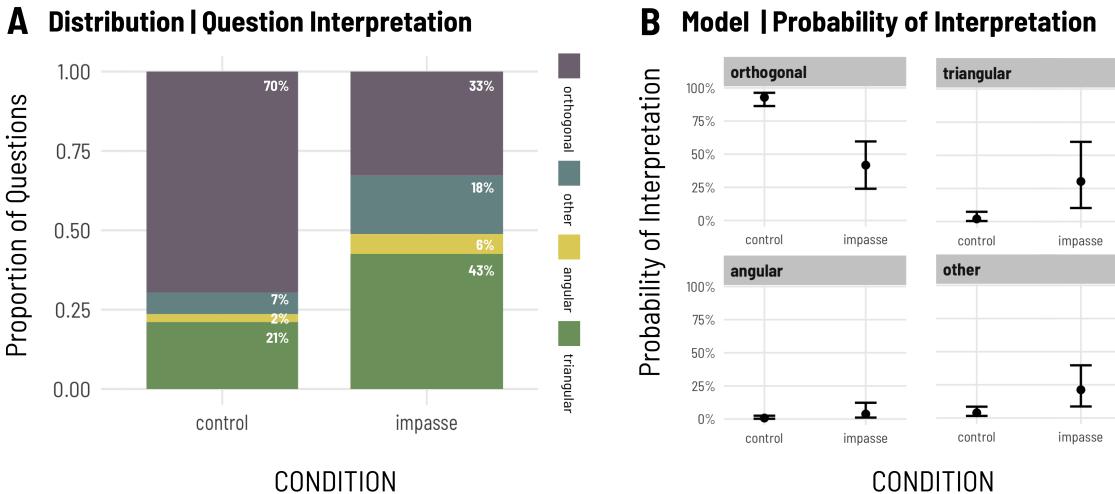


Figure 3.7. Study 3A – Interpretation. [A] A proportional bar chart shows the impasse condition increases the proportion of non-orthogonal interpretations. [B] The model predicts a significantly higher probability of *other*, *angular* and *triangular* responses in the impasse condition.

the odds of transitional interpretations. Across the entire task participants in the impasse condition were 12 times as likely to offer an *other* rather than orthogonal response compared with those in the control condition ($e^{\beta_1} = 12.13$, 95% CI [6.29, 25.24], $pd = 100\%$). Participants in the impasse condition were also 12 times more likely to offer an *angular* (vs) orthogonal response compared with those in the control ($e^{\beta_1} = 11.48$, 95% CI [3.95, 37.67], $pd = 100\%$), and 34 times more likely to offer a *triangular* rather than orthogonal response compared with those in the control condition ($e^{\beta_1} = 33.90$, 95% CI [6.22, 211.18], $pd = 100\%$). Raw data are visualized in Figure 3.7 [A], model predictions in Figure 3.7 [B] and parameter estimates and model specification are detailed in Appendix B.1.2.

3.3.3 Study 3A Discussion

In Study 3A we find evidence in support of the impasse hypothesis: that intentionally constructing an obstacle to the most likely misinterpretation of a novel graphical formalism can significantly improve the accuracy of its interpretation. Participants in our impasse condition

were both more likely to produce correct-triangular responses, but also technically incorrect responses that were *not* orthogonal: combinations of points that were either *angular* (following the diagonal or horizontal gridlines), or blank or some other unidentified alternative. Thus, it seems that imposing a structure upon a graph reading task designed to invoke a state of mental impasse helps some respondents correctly restructure their understanding to reach a *completely correct* solution, and in other cases, simply restructure *away* from the *most incorrect* solution.

These intermediary response types offer support for an important distinction Ohlsson made in his information-processing account of insight: that a *series* of insights may be needed before a complex problem is solved (1992). An individual insight does not necessarily yield a complete, correct solution, and a correct solution may never be reached. To Ohlsson, insight by definition necessitates a state of mental impasse, but mental impasse does not guarantee a moment of insight.

3.4 Study 3A: Online Replication

During the Covid-19 pandemic we performed an online replication of Study 3A to validate the use of the Interval Graph Comprehension Task when performed in an remote, asynchronous modality. We added additional attention check questions to the stimulus web-application, as well as browser interaction tracking allowing us to exclude participants who leave the browser window during the study. The stimulus web-application required a minimum screen resolution of 1125px X 680px, the Chrome web-browser, and either an external mouse or trackpad (i.e. the study could not be completed on a mobile or touchscreen device). To evaluate these enhancements we recruited 107 undergraduate students at UC San Diego to participate (online, asynchronously) in exchange for course credit. Thirty-six individuals (34% of total recruitment) were excluded for either failing attention-check questions, or violating browser interactions (such as leaving or resizing the experiment window), yielding 71 participants for analysis (28 % male, 68 % female, 4 % other; age: 18 - 27 years). Participants followed the same procedure as Study

3A, but did so online, asynchronously, using their own (laptop or desktop) computer. The same task and analysis parameters were used as described in Section 3.3.1. The distribution of total score is visualized in Figure 3.8. Overall, we see the same pattern of behaviour as Study 3A (Figure 3.4). **Consistent with our hypotheses and results from Study 3A, posing a mental impasse had a statistically significant, positive effect on the probability of a correct response across questions, and increased the proportion of *unknown/uncertain, angular and triangular responses, relative to incorrect orthogonal responses*.** Full model specifications and results are detailed in Appendix B.2. These results provide converging evidence that posing a mental impasse has a reliable, positive impact on some readers' ability to correctly interpret the coordinate system. It also verifies that although the interval graph comprehension task is challenging, reliable results can be obtained via online experimentation with more stringent exclusion criteria driven by additional attention checks and browser interaction logging.

STUDY 3A (Online Replication) | Distribution of Total Score

Impasse condition yields greater variance and more high scores

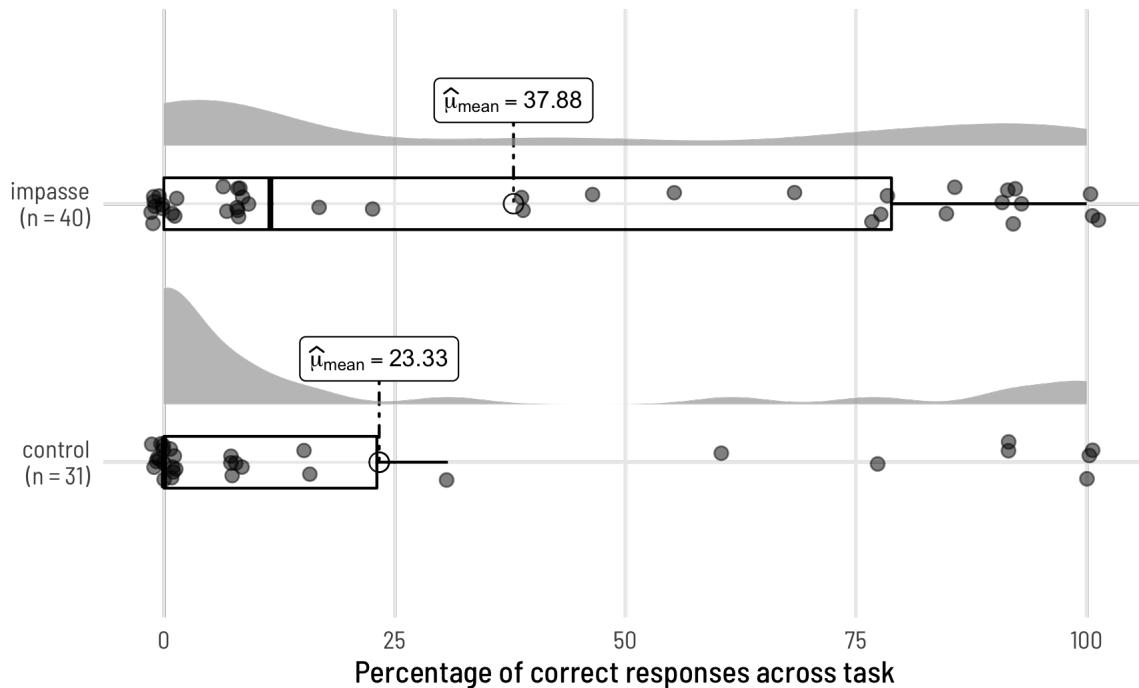


Figure 3.8. Study 3A Online Replication — Distribution of Total Score. Consistent with the synchronous in-person results of Study 3A, the impasse condition yielded higher scores on the Interval Graph Comprehension task.

3.5 Study 3B: Implicit (vs) Explicit Scaffolding

In Study 3A we verified that imposing an impasse-structure targeted at a reader's most likely misconception can improve discoverability of a novel graphical formalism. In Study 2 we found evidence that providing explicit guidance (in the form of text or image instructions) also facilitates discovery. Our goal in Study 3B is to explore these two forms of scaffolding: one *implicit* and the other *explicit*, in combination. Which is more effective? Do they have an additive, or rather, interacting effect?

Our hypotheses for this investigation are motivated by the finding (consistently across Studies 1-3A) that some readers persist in incorrectly interpreting the TM graph, even when provided with scaffolding. We can interpret this in two ways. First that the effect of prior

knowledge of common graphical formalisms (i.e. the cartesian coordinate system) and learned behaviour for graph reading (i.e. the procedures for extracting information from a cartesian coordinate system) are extraordinarily difficult to overcome. Secondly, that it is often not clear to the reader when they have come to a mistaken interpretation. While providing explicit instructions may help to overcome the first issue, it does not address the second. We believe that an impasse structure, however, does address this, by offering the reader implicit feedback that their interpretation is incorrect (when there is no available answer to a question). In the impasse-structure condition a greater burden is placed on the reader to discover the rules of the new representational system, but it should at least be clear to them that they *need to do so*. Alternatively, in the image-based explicit scaffold conditions, the reader has a lesser burden on discovering the rules (i.e. they are made explicit), but these instructions can be ignored, if the reader does not believe they are needed (i.e. they do not realize they are misreading the graph). For this reason we expect that while both explicit and implicit forms scaffolding will result in more readers accurately interpreting the graph, they will yield different effects on transitional interpretations.

Specifically, we hypothesize that:

(H1) In replication of Studies 2 and 3, both *implicit* (impasse) scaffold and *explicit* (image-based) scaffolds will improve response ACCURACY relative to a *non-image* and *non-impasse* control conditions.

(H2) The *impasse* scaffold will yield more transitional INTERPRETATIONS (*other*, *angular*) than the *explicit* scaffolds, which will instead be more effective at producing correct, *triangular* interpretations

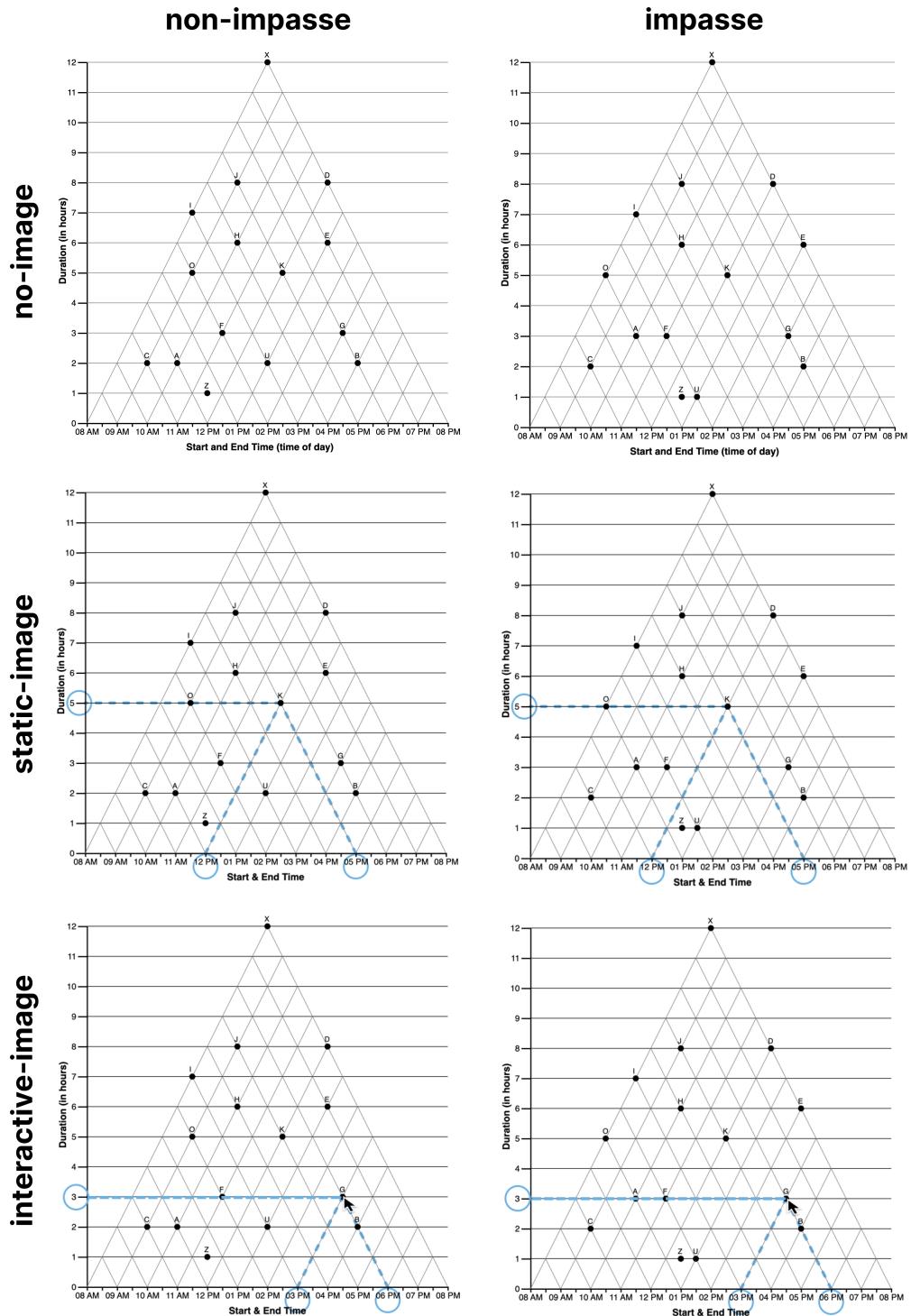


Figure 3.9. Study 3B — Experimental Conditions. The first column shows the structure of the graph for the *non-impasse* level of the **implicit** scaffold factor, while the second column shows structure of the *impasse* level. The top row shows the view of the *no-image* (control) level of the **explicit** scaffold factor. The second row shows the view for the *static-image* level, and the bottom row the view of the *interactive-image* level. Note that in the *interactive-image*, the blue lines are contextual and appear over-top each datapoint when the mouse cursor is hovered above them.

3.5.1 Methods

3.5.1.1 Participants

We recruited 373 undergraduate students at UC San Diego to participate (in person) in exchange for course credit (gender: 39 % male, 60 % female, 2 % other; age: 18 - 66 years).

3.5.1.2 Design

The experiment was defined by a multilevel factorial structure with 2 fixed and two random factors:

(F1) **explicit scaffold** (between-subjects) @ (c = 3) levels: *no-image* [control], *static-image*,

interactive-image

(F2) **implicit scaffold** (between-subjects) @ (c = 2) levels : *non-impasse* [control], *impasse*

(R1) **question** (within-subjects) @ (q = 13) levels

(R2) **participant** @ (n = 373) levels

The two fixed factors were fully crossed, yielding six distinct scaffold conditions: *no-image* |*non-impasse*, *static-image* |*non-impasse*, *interactive-image* |*non-impasse*, *no-image* |*impasse*, *static-image* |*impasse*, *interactive-image* |*impasse*. Participants were nested within condition, and questions were fully crossed with condition. Thus, each participant was randomly assigned to one of the six (factorial) conditions, in which they completed all questions.

3.5.1.3 Materials & Procedure

Interval Graph Comprehension Task

The Interval Graph Comprehension Task from Study 3A was used (with identical dataset, questions and scoring strategy) as described in Section 3.3.1.3. The experimental conditions were defined by the two most effective scaffolding techniques form Study 2 (the static and interactive images) fully crossed with with the impasse-structure from Study 3A (i.e. in a

factorial design). Figure 3.9 depicts the stimulus graph displayed in the first question for each condition. Participants followed the same procedure as Study 3A as described in Section 4.3.1.4.

3.5.1.4 Analysis

Response Accuracy

To test hypotheses related to response ACCURACY we fit mixed logistic regression models in R using the `lme4` package. For contrast coding, the default treatment (dummy) coding scheme was used with the following reference categories:

- response variable ACCURACY : level *0:incorrect* as reference
- predictor factor **implicit scaffold**: level *non-impasse (control)* as reference
- predictor factor **explicit scaffold**: level *no-image (control)* as reference

To determine a final model we first defined the maximal random effects structure theoretically justified by the study design (random intercepts for questions and subjects). We then fit the most complex model indicated by the study design, including (main) fixed effects for **implicit** and **explicit** scaffold as well as their interaction term, and used a likelihood ratio test to determine if this model was superior to a simpler model including fixed main effect only. Statistical significance of each predictor in the superior model was determined via Wald Chi-Square tests, and all reported p-values are for non-directional tests with a decision threshold $\alpha = 0.05$.

Response Interpretation

To test hypotheses related to response INTERPRETATION we fit Bayesian mixed multinomial regression models in R using the `brms` package. We used the same execution parameters as defined for Study 3A (Section 3.3.1.5) The default treatment-coding scheme was used with the following reference categories:

- response variable INTERPRETATION : level *orthogonal* as reference
- predictor factor **implicit scaffold**: level *non-impasse (control)* as reference

- predictor factor **explicit scaffold**: level *no-image (control)* as reference

To determine a final model we first defined the maximal random effects structure theoretically justified by the study design (random intercepts for questions and subjects). We then fit models including (main) fixed effects for **implicit and explicit scaffold**, and a more complex model including their interaction term, and used a Bayes Factor model comparison to determine if there was sufficient evidence in support of the more complex model. In addition to the informative priors described in Section 3.3.1.5, we set a direction-neutral prior on the interaction term : normal ($\mu = 0, SD = 2.42$).

3.5.2 Results

3.5.2.1 Overall Accuracy

To explore the effect of **explicit** and **implicit** scaffolding on TM graph interpretation, we start by describing the distribution of TOTAL SCORE, operationalized as the percentage of correct responses (derived from the ACCURACY measure) of the 13 interpretation-discriminant questions of Interval Graph Comprehension Task. Across all conditions, TOTAL SCORE ranged from 0 to 100 with a mean of a mean of 61%. Note this is nearly twice the average score from Study 3A: a substantial improvement! In Figure 3.11 we see that total scores steadily increased across explicit scaffold conditions, such that no-image (control) < static-image < interactive-image. Further, within each explicit scaffold factor, the addition of the impasse manipulation further improved TOTAL SCORE. The difference between the lowest scoring condition (*no-image |non-impasse*) ($M = 21\%, SD = 0.37, n = 62$) and the highest scoring condition (*interactive-image |impasse*) ($M = 84\%, SD = 0.24, n = 61$) is substantial. This pattern of results is indicative of a likely main effect of both explicit and implicit scaffold factors. Notice that across the interactive-image conditions (*impasse* vs. *non-impasse*) the mean score and variance are very similar—nearly at ceiling—compared with a salient difference between *impasse* and non-impasse in absence of an explicit scaffold (*no-image*, top facet). This suggests there may also be an interaction effect between implicit and explicit scaffolds.

STUDY 3B | Distribution of Total Score

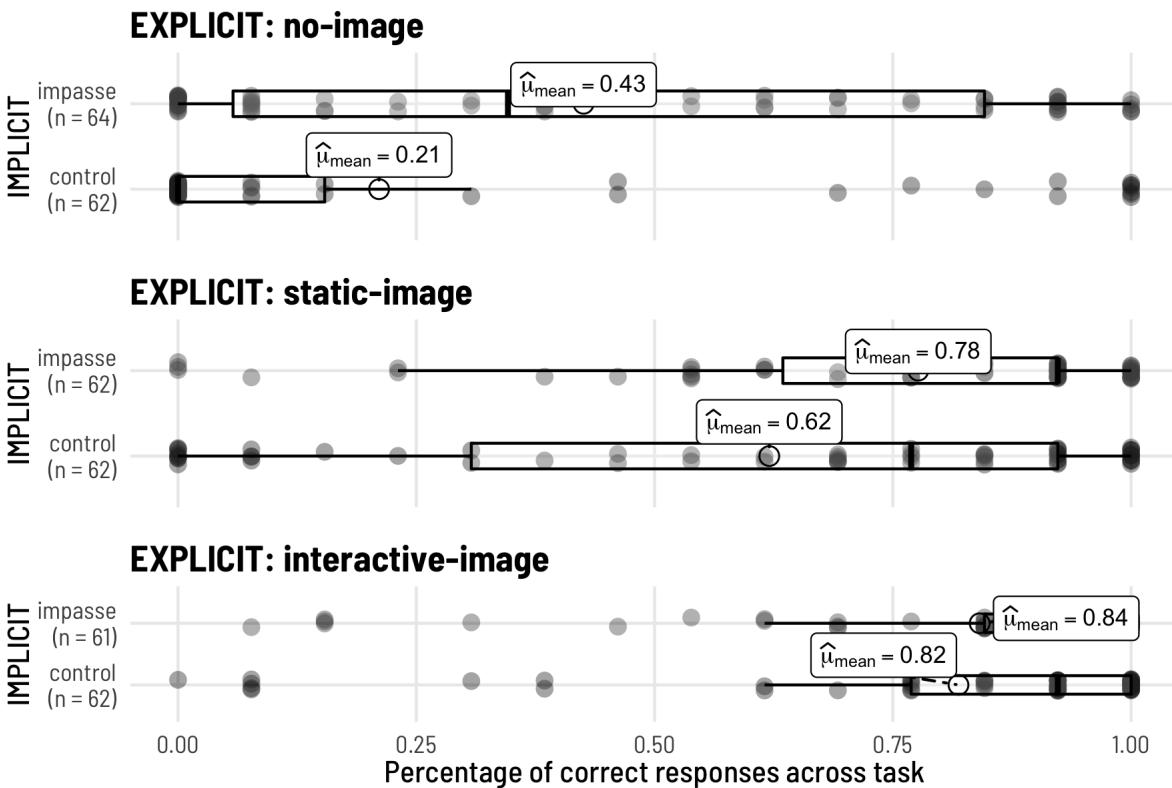


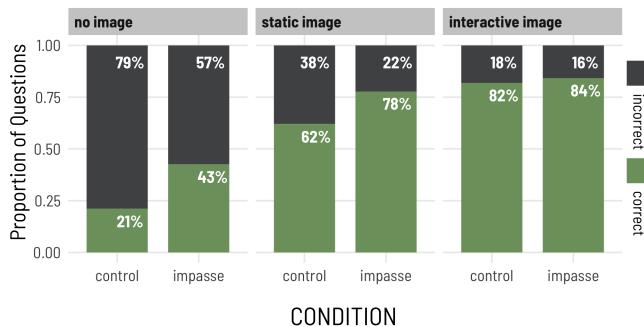
Figure 3.10. Study 3B — Distribution of Total Score. Consistent with the results of Study 2 and Study 3A, it appears that the (explicit) static-image and interactive-image conditions yield higher total scores than the no-scaffold control condition. Further, within each explicit scaffold factor, the impasse condition yields higher scores. This pattern of results is suggestive of main effects and possible interaction between explicit and implicit scaffold factors.

3.5.2.2 Accuracy

To quantify the effect of **implicit** and **explicit** scaffolding techniques on ACCURACY, we fit a mixed effects logistic regression model with random intercepts for subjects and questions, and **implicit**, **explicit** and their interaction term as fixed effects. A likelihood ratio test indicates that a model including both fixed effects and their interaction term is a significantly better fit to the data than a simpler main-effects model ($\chi^2(6, 8) = 8.20, p < 0.05$). The explanatory power of the final model is substantial (*conditionalR*² = 0.83) and the part related to the fixed effects (*marginalR*²) explains 32% of variance.

STUDY 3B | ACCURACY

A Distribution | Question Accuracy



B Model | Probability of Correct Response

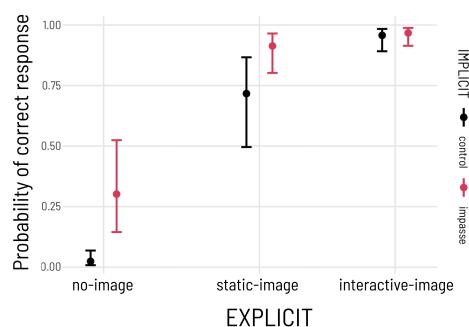


Figure 3.11. Study 3B – Accuracy. [A] A proportional bar chart of raw data shows the relative percentage of correct (green) responses steadily increases across the explicit scaffold conditions, reaching near ceiling in the *interactive-image* conditions. [B] The model predicts a significantly higher probability of correct response for each *impasse* condition, across *no-image* and *static-image* explicit scaffolds.

Consistent with our (H1) hypothesis, both explicit and implicit scaffolding improves response ACCURACY. Wald Chi-Square tests revealed both significant main effects for **implicit** ($\chi^2(1) = 19.3, p < 0.001$) and **explicit** ($\chi^2(2) = 95.2, p < 0.001$) scaffolds, as well as a significant interaction ($\chi^2(2) = 8.1, p < 0.05$). The regression coefficients indicate that across **explicit** scaffolds, the *impasse* manipulation increases the odds of a correct response by a factor of 17.5, ($e^{\beta_1} = 17.5, SE = 11.5, p < 0.001$). The effect of **explicit** scaffolds was much larger, however. Across both **implicit** scaffold conditions, the *static-image* increases odds of a correct response by a factor of 103, ($e^{\beta_1} = 103, SE = 68.7, 381], p < 0.001$) and the *interactive-image* increases odds of a correct response by a factor of 910, ($e^{\beta_1} = 910, SE = 643, p < 0.001$).

Regarding the interaction effect, post-hoc paired comparisons (with Tukey method correction) reveal that across each the *no-image* control and *static-image* conditions, posing a mental *impasse* significantly improved performance over the *non-impasse* control. The source of the interaction effect is driven by response ACCURACY reaching ceiling across both *non-impasse* and *impasse* conditions ($OR = 0.76, z = -0.43, p = 0.99$). That is to say, the explicit *interactive-image* scaffold is so effective, very few participants are in need of additional *impasse* scaffolding.

This is in stark contrast with the *no-image* explicit scaffold conditions, where the addition of the mental *impasse* significantly increased the probability of a correct response ($OR = 0.06, z = -4.39, p < 0.0001$). Raw data are visualized in Figure 3.11 [A], model predictions in Figure 3.11 [B] and parameter estimates and model specification are detailed in Appendix B.3.1.

3.5.2.3 Interpretation

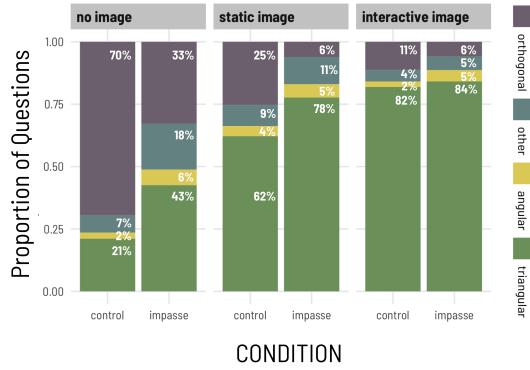
To test the hypothesis that posing a mental impasse will yield more transitional interpretations of the interval-coordinate system than occur with explicit instructions, we quantify the effect of **implicit** and **explicit** scaffolding on INTERPRETATION by fitting a Bayesian mixed multinomial regression model with random intercepts for subjects and questions. A Bayes Factor analysis comparing a main effects only model vs a more complex model including the interaction term between **implicit and explicit scaffolds** indicates extreme evidence in favour of the *simpler* main effects only model ($BF = 4.04 \text{ e+122}$). In Figure 3.12 [A] we see first that the proportion of correct *triangular* responses steadily increases across impasse and explicit scaffold conditions, lowest in the *no-image |non-impasse* condition, and nearing ceiling in the *interactive-image |impasse* condition. Across both *no-image* and *static-image* explicit scaffolds, the *impasse* yields a greater proportion of *other* and *angular* type responses. But these are similarly minimized in the presence of the *interactive-image* scaffold.

Consistent with our findings for question ACCURACY, we see similarly increasing probability of triangular responses across the explicit scaffold factors such that *no-image* < *static-image* < *interactive-image*. Within each explicit scaffold, we see that the *impasse* condition yields more *angular* and *other* responses relative to *non-impasse*, consistent with our findings in Study 3A. Much like with response ACCURACY, we see near ceiling performance across both non-impasse and impasse conditions with the *interactive-image* scaffold.

The model parameter estimates (see Appendix B.3.2) indicate reliable evidence for main effects of both **implicit** and **explicit** scaffolds. **Consistent with our (H2) hypothesis, the impasse factor yielded a greater increase in odds for other and angular responses (compared**

STUDY 3B | INTERPRETATION

A Distribution | Question Interpretation



B Model Prediction | Probability of Interpretation

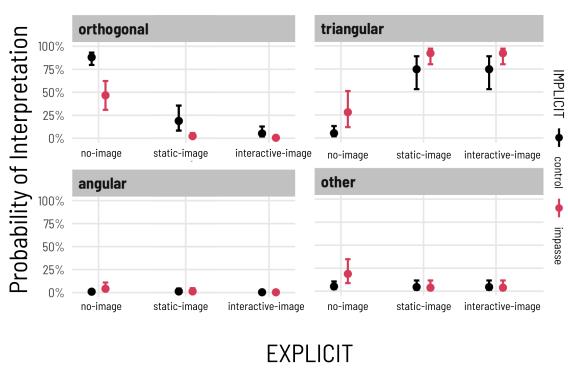


Figure 3.12. Study 3B – Interpretation. [A] A proportional bar chart shows an additive effect of implicit and explicit scaffolding, reaching near ceiling in the presence of the *interactive image* [B] The model predicts a significantly higher probability of *triangular* responses, and similarly probabilities for *other*, and *angular* responses across impasse and explicit scaffolds.

to the *static* and *interactive* image conditions. However, both **explicit** scaffolds had a much stronger effect on the odds of a correct triangular response. Raw data are visualized in Figure 3.12 [A], model predictions in Figure 3.12 [B] and parameter estimates and model specification are detailed in Appendix B.1.2.

3.5.3 Study 3B Discussion

In Study 3B we've found a pattern of results consistent with our findings in Study 3A and its online replication: in the absence of an explicit scaffold, invoking an mental impasse improves interpretation accuracy. We also replicated the results of Study 2 with respect to image-based explicit scaffolds, which dramatically improved accuracy relative to a *no-image* control. Together, we found these scaffolding techniques have an additive effect. Both static and interactive image scaffolds dramatically increased the odds of a correct response, which were pushed yet higher in the presence of an impasse structure, until they reach near-ceiling in with the presence of an *interactive-image* scaffold. We also found evidence in favour of our H2 hypothesis: the implicit factor yields more transitional (*other*, *angular*) interpretations than the explicit factor, which in turns yields more *triangular* responses. From these results we can

conclude that both explicit and implicit scaffolding are effective at supporting discovery of the TM graph, and that they work well *in combination*. A designer may wish to offer both kinds of scaffolds when introducing a novel graphical form in an a scholarly paper, for example. One important question that remains, however, is whether the additive effect of the two scaffolds is a result of one individual requiring *both* techniques in order to restructure their understanding, or alternatively, if they each appeal to different individuals, but taken together, result in the correct restructuring of a greater proportion of participants.

3.6 Study 3C: The Role of Working Memory

Although we've found evidence of reliable, positive effects of both explicit (Study 2 Study 3B) and implicit (Study 3A, Study 3B) scaffolding on interval graph discovery, we've also seen considerable variance across participants. Without any intervention, some individuals in the non-scaffolded (control) conditions are still able to reach a correct-triangular interpretation, and despite substantial scaffolding (even a combination of interactive-image and impasse structure), some individuals persist in an incorrect cartesian interpretation. In Study 3C we explore one potential source of these individual differences: **working memory capacity**.

There is a long history of research examining the role of working memory capacity (WMC) in both diagram/visualization comprehension and insight problem solving. Individual differences in visual-spatial tasks including diagrammatic reasoning have been consistently connected to the visual-spatial working memory system (Just and Carpenter, 1992; Sims and Hegarty, 1997). In the context of visualization research, WMC has been considered a limiting factor in one's ability to effectively extract information from, or perform subsequent operations on, visualized information. In early visualization research the conventional wisdom was that visualizations (including graphs) were effective because they 'offloaded' some of the burden of cognition onto visual perception; an acknowledgement of the capacity limits in cognitive processing. Lohse (1997) noted, however that adding a visualization to a task did not always

improve performance. He evaluated the effect of WMC on accuracy of decision making with information presented in a graph, and found that it was only at high levels of task complexity (i.e. only when the working memory system was ostensibly highly taxed) that individual differences in WMC explained performance. Predictably, it was only participants with lower WM capacity whose performance improved with the addition of a better designed graphical aid. Providing a better visualization helped those with low WM match the accuracy of those with high WM, who did not *need* the better visualization. This line of evidence indicates that individual differences in WMC are relevant to performance on graph and visualization tasks, and further, that changes in the features of a visualization may not universally change performance, but rather help to ameliorate specific deficits (for specific readers), such as working memory capacity.

Similarly, the role of working memory has also been explored with respect to insight problem solving. One dimension along which accounts of successful problem solving differ is the extent to which individual differences are best explained by adoption of particular strategies, versus constructs related to general intelligence such as working memory capacity. Murray & Byrne have claimed that a crucial property of insight problems is that they require solvers to simultaneously consider alternative possibilities, thus taxing working memory (Byrne and Murray, 2005). But solvers also need to effectively switch attention between alternatives in order to reach an appropriate solution. In this way, the insight literature carefully differentiates between working memory capacity, and the allocation of working memory via attentional switching. In an empirical investigation Byrne & Murray found that high WMC and high ability on attentional switching predicted insight problem solving performance, but measures of focused and sustained attention did not (2005), indicating both that the relationship between these constructs, their measures and different kinds of insight problems, is likely nuanced.

Working Memory Capacity and Graph Discovery

With respect to the Interval Graph Comprehension Task, and the broader phenomenon of discovering the rules of a representational system that we are construing in this work as a type of

insight problem, we have reason to suspect that high working memory capacity (WMC) might facilitate the kind of problem restructuring required to transition from a cartesian-scatterplot to triangular-interval interpretation of the TM Graph. However, it is also plausible that high working memory capacity might facilitate the execution of suboptimal strategies. For example, it is trivial to read the *end time* property from the TM Graph *if* you use its triangular coordinates. If you use cartesian coordinates, extracting the end time property is computationally costly, requiring visual search over multiple potential start times, reading duration from the y-axis, horizontal project to derive the end-time, and then projection to the x-axis (see Figure 3.13). It is possible that an individual with high WMC might persist in this costly operation until they derive an available answer, while an individual with low WMC might find the operation unfeasible, they might try and fail and give an orthogonal-like but also incorrect response (show examples) or alternatively the high cost of this computation could pose an impasse for those with low WMC, prompting them to restructure to a triangular interpretation.

Thus, our exploration of the relationship between working memory capacity and performance on the Interval Graph Comprehension Task is *exploratory*. We offer no directional hypotheses, but rather, ask the following research questions:

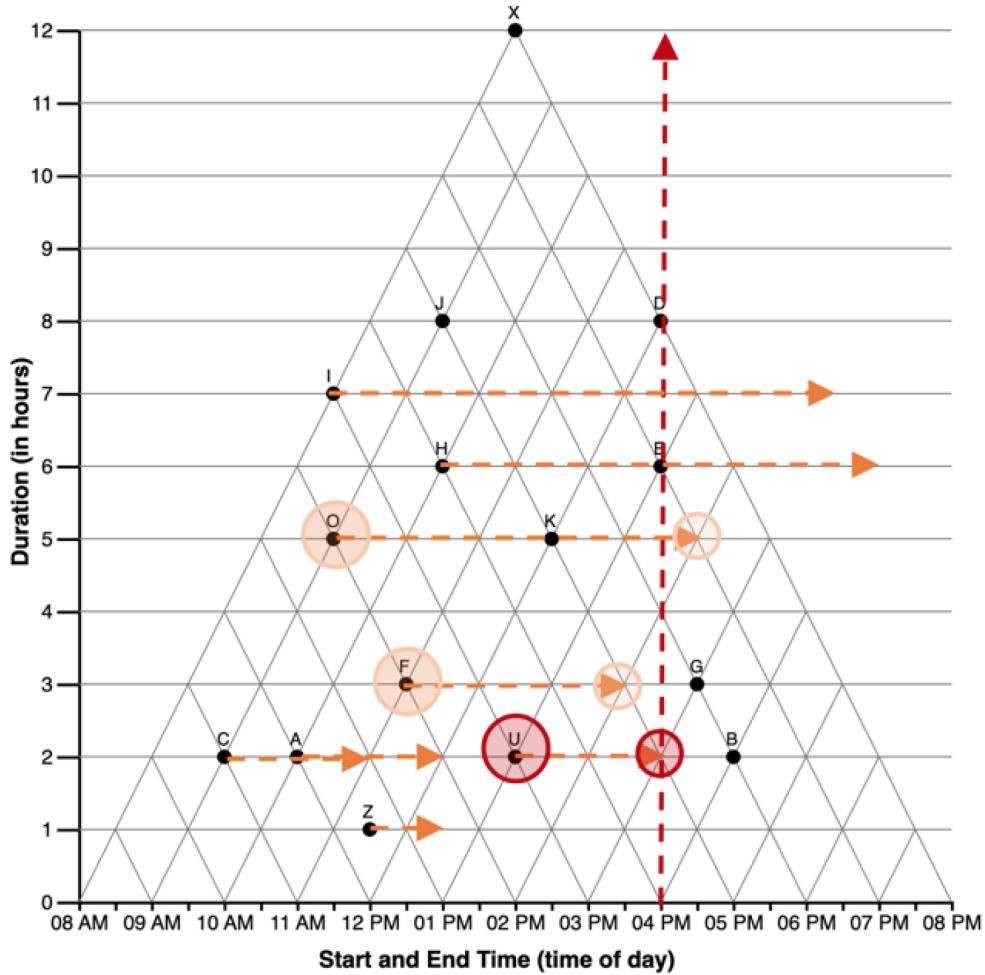
- (R1) Does WMC explain variance in accuracy or interpretation in the Interval Graph Comprehension Task?
- (R2) Does application of an impasse-structure differentially affect accuracy and interpretation in the presence of an *impasse-structure* scaffold?

3.6.1 Methods

3.6.1.1 Participants

We recruited 200 undergraduate students at UC San Diego to participate (online, asynchronously) in exchange for course credit. 67 participants were excluded for failing attention-check questions, browser violations or failure to complete the working memory task, yielding 133 participants for analysis (gender: 32 % male, 67 % female, 2% other; age: 18 - 31 years).

(non-impasse) control



Q4: What event(s) end at 4pm?

Figure 3.13. Study 3C – Costly End Time Calculation. Answering end-time property questions using an orthogonal (cartesian) interpretation incurs high cost on working memory (without the ability to draw or otherwise make marks on the graph). Using the orthogonal interpretation, the reader: (1) locates 4pm on the x-axis, (2) draws an orthogonal projection upwards, (3) locates data points to the left of the projection (4) *for each* data point, reads the duration off the y-axis, and projects rightward until calculating an end time, (5) determines which if any points end at the orthogonal projection with 4pm. The (incorrect) orthogonal answer is point **U**. Two common answers were points **F** and **O**, both within half an hour of the orthogonal projection. The correct (triangular) answer is point **B**. For this question, correctly reading the graph should place less load on working memory than incorrectly reading the graph.

3.6.1.2 Design

The experiment employed a multilevel design structure with 1 fixed and 2 random factors:

(F1) **implicit scaffold** (between-subjects) @ 2 levels : *none* [control], *impasse*

(R1) **question** (within-subjects) @ ($q = 13$) levels

(R2) **participant** @ ($n = 133$) levels

Participants were nested within **implicit** scaffold condition, and questions were fully crossed with condition. Thus, each participant was randomly assigned to one condition, in which they completed all the questions. Note that the dependent measure Working Memory Capacity (**WMC**) was added to each statistical model and treated as a fixed effect.

3.6.1.3 Materials

Interval Graph Comprehension Task

Participants were randomly assigned to an **implicit scaffold** condition (*non-impasse* or *impasse*), in order to complete the Interval Graph Comprehension Task (as described in Study 3A, Section 3.3.1.3).

OSPAK Working Memory Task

Working memory is a complex psychological construct and thus methods for measuring its properties (including capacity) are both varied and hotly debated. One popular method of measuring working memory capacity (WMC) in the contemporary cognitive literature is the Operation Span (OSPAK) task (Oswald, McAbee, Redick, and Hambrick, 2015). In the OSPA, a participant is asked to solve a series of simple arithmetic problems while simultaneously remembering a set sequentially-presented stimuli. The classic OSPA task uses a series of letters or words as the memory stimuli (Oswald, McAbee, Redick, and Hambrick, 2015). In theory, this requires that participants hold information in mind (the series of stimuli) for a short period of time, while simultaneously performing the information processing required to execute the

arithmetic operations, which ostensibly involve different cognitive mechanisms. Converging evidence has shown high test-retest reliability for the OSPAN, as well as strong correlation with other WMC assessments including symmetry and reading span tasks (Conway, Kane, Bunting, Hambrick, Wilhelm, and Engle, 2005), and sufficient sensitivity to detect differences between participants who perform as novices (vs) experts on other higher order cognitive tasks.

To maximize validity of the OSPAN as a measure of WMC however, it is important to ensure that participants engage in the task *as designed*. Hicks & colleagues (2016) recently developed a version of the task which replaces the linguistic stimuli (which are more easily rehearsed, sub-vocalized or written down) with simple icons (e.g. image of car, house, bird, etc), which they demonstrated were more reliable for data collection in (unsupervised) asynchronous online environments.

In Study 3C we use a web-based version of the Hicks & colleagues (2016) pictorial-OSPAK task developed for via Qualtrics by Castro & colleagues (2021). In this task a trial proceeds by showing participants a sequence of (either 4, 5 or 6) simple images, which they are instructed to remember in order. After the last image, they are presented with a simple math equation (e.g. $(4 \times 2) + 1 = 5$), and asked to click a button to indicate if the equation is TRUE or FALSE. They are then presented with an array of pictures, and asked to click on the images they remember seeing, in the order they recall them being presented. After receiving instructions and two practice trials, participants complete six experimental trials (with 4,5, and 6-image spans) presented in random order.

Measures and Scoring The OSPAN task yields two raw scores: one for accuracy in remembering the sequence of images (sequence score), and one for accuracy in responding to the math problems (math score). Sequence scores are assigned by calculating the number of images the participants correctly report in order (for example, if the participant selects the correct 4 images but only 1 is in the correct order, they receive 1 out of 4 points). A single weighted score is then calculated by multiplying each participant's proportion of correct math problems, by their

total sequence score. This approach ensures that participants who neglect one aspect of the task (sequence or math) in favour of prioritizing the other do not receive inappropriately high scores. To achieve a perfect weighted score, a participant must remember all sequences of items, and correctly answer all math problems. Finally, following convention in the visualization individual differences literature, we performed a median-split on the sample, thus dividing participants into two groups on a covariate factor we refer to as **WMC** (two levels: *low-memory, high-memory*)⁵.

3.6.1.4 Procedure

Participants completed the study in online, asynchronously, using their own (laptop or desktop) computer. After agreeing to an IRB-approved informed consent, participants were randomly assigned to an **implicit scaffold** condition and presented with task instructions. They then completed the Interval Graph Comprehension Task. Upon completion of the graph comprehension task, participants were re-directed to a Qualtrics survey which administered the OSPAN Working Memory task. After the OSPAN task, participants were presented with a demographic questionnaire, and final debriefing text.

3.6.1.5 Analysis

Response Accuracy

To test hypotheses related to response ACCURACY we fit mixed logistic regression models in R using the `lme4` package. As in Study 3A, the default treatment (dummy) coding scheme was used with the following reference categories:

- response variable ACCURACY : level *0:incorrect* as reference
- predictor factor **implicit scaffold**: level *non-impasse (control)* as reference
- covariate factor **WMC**: level *low-memory* as reference

⁵We also fit a model with equivalent random effects structure using a mean-centred OSPAN weighted score as a continuous variable, and found the same pattern of results yielding the same statistical inferences. Here we report results for the 2-group median split to facilitate interpretation of the interaction term.

To determine a final model we first defined the maximal random effects structure theoretically justified by the study design (random intercepts for questions and subjects). We then fit a model with **implicit scaffold** and WMC as simple fixed effects and used a likelihood ratio test to determine superior fit between this and a more complex model including the **implicit : WMC** interaction term. Statistical significance of each predictor in the final model was determined via Wald Chi-Square tests, and all reported p-values are for non-directional tests with a decision threshold $\alpha = 0.05$.

Response Interpretation

To test hypotheses related to response INTERPRETATION we fit Bayesian mixed multinomial regression models in R using the `brms` package. We used the same model execution parameters as defined for Study 3A (Section 3.3.1.5) The default treatment-coding scheme was used with the following reference categories:

- response variable **interpretation** : level *orthogonal* as reference
- predictor factor **implicit scaffold**: level *non-impasse (control)* as reference
- predictor factor **WMC**: level *low-memory* as reference

To determine a final model we first defined the maximal random effects structure theoretically justified by the study design (random intercepts for questions and subjects). We then fit a model with both **implicit scaffold**, **WMC** and their interaction term as fixed effects and used a Bayes Factor model comparison to determine if there was sufficient evidence in support of the predictor model over the random effects only model. In addition to the informative priors described in Section 3.3.1.5, we set a direction-neutral prior on the interaction term : normal ($\mu = 0, SD = 2.42$).

3.6.2 Results

3.6.2.1 Overall Accuracy

The distribution of TOTAL SCORE follows the same pattern of behaviour observed in Study 3A and its online replication: the distribution is bimodal (Figure 3.14) . Total scores were higher in the *impasse* condition ($M = 35\%$, $SD = 41\%$, $n = 65$) than *non-impasse* control condition ($M = 15\%$, $SD = 30\%$, $n = 68$), implying a likely main effect of **implicit** scaffold. Comparing total scores across the median split in **WMC** (*high* (vs) *low* working memory capacity), we see that readers with *high WMC* ($M = 31\%$, $SD = 41\%$, $n = 66$) performed better than readers with *low WMC* ($M = 19\%$, $SD = 31\%$, $n = 67$). Although readers in the *impasse* condition performed consistently better than those in the *non-impasse* control, the effect is particularly pronounced for readers with *high WMC*, implying a potential interaction between **implicit** scaffold and working memory capacity.

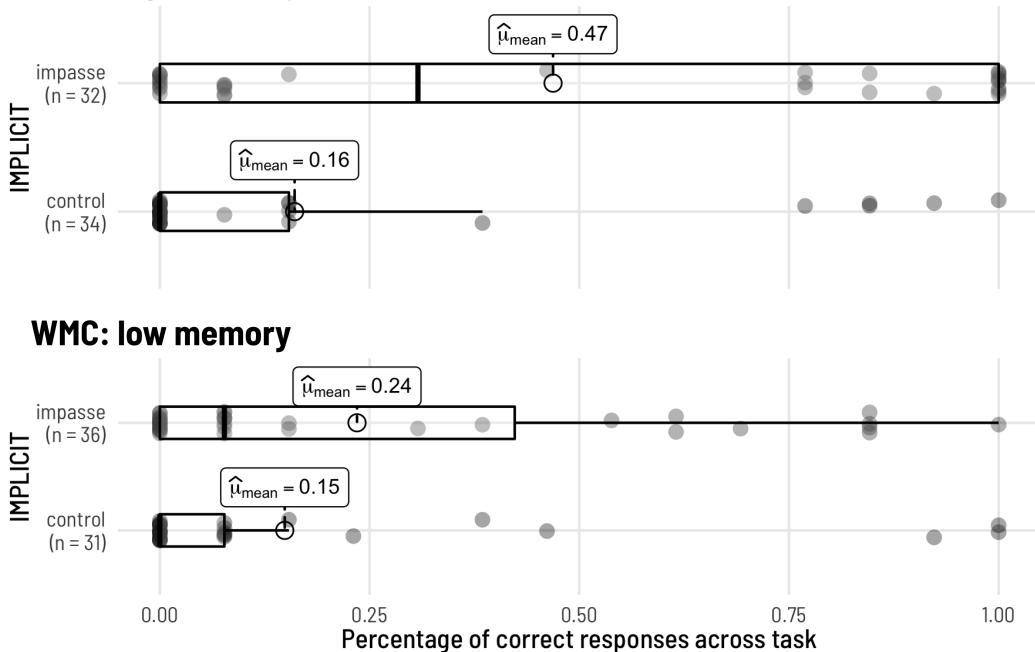
3.6.2.2 Accuracy: Does working memory affect response accuracy?

To quantify the effect of working memory capacity on ACCURACY, we fit a mixed effects logistic regression model with random intercepts for subjects and questions, and **implicit** scaffold, **WMC** and their interaction term as fixed effects. A likelihood ratio test indicates that a model including the the interaction term explains significantly more variance in accuracy than a main-effects only model ($\chi^2(5, 6) = 5.04, p < 0.05$). The explanatory power of the entire model is substantial (*conditional R²* = 0.92) and the part related to the fixed effects (*marginal R²*) explains 18% of variance.

Wald Chi-Square tests revealed no significant main effects, but rather a significant interaction between **implicit** scaffold and **WMC**. Post-hoc paired comparisons (with Tukey method correction) indicate that for participants with low working memory capacity, performance was comparably low, regardless of which experimental condition was randomly assigned ($OR = 0.1, SE = 0.15, z = -1.47, p = 0.46$). Similarly, for those who were assigned to the *non-impasse* control condition, working memory capacity **did not** significantly influence performance ($OR =$

STUDY 3C | Distribution of Total Score

WMC: high memory



WMC: low memory

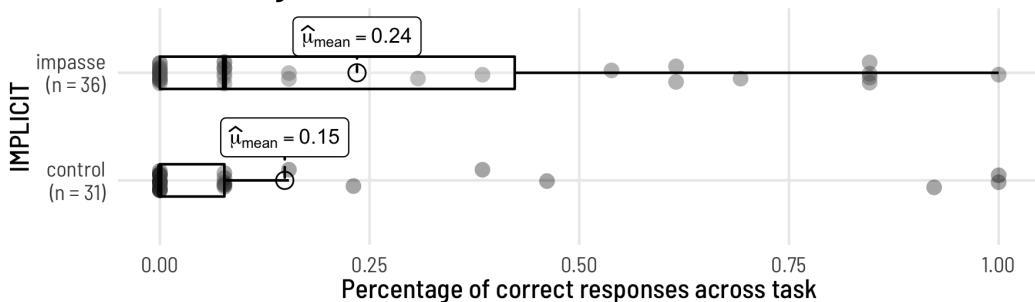


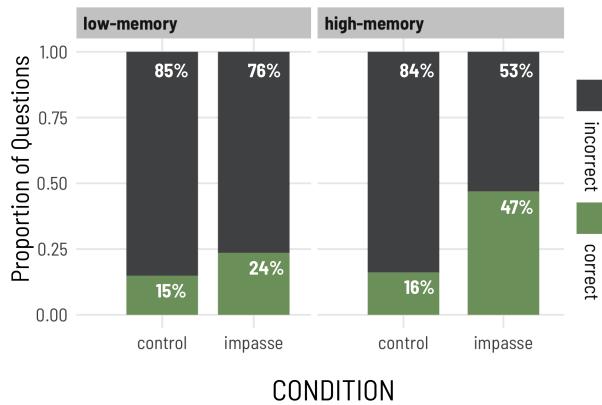
Figure 3.14. Study 3C – Distribution of Total Score. The mean score in the *impasse* condition for participants with *high working memory capacity* (top facet of graph) is nearly double that of individuals with *low WMC* randomly assigned to the *impasse* condition.

1.95, SE = 2.80, z = 0.47 p = 0.97). For those assigned to the *impasse* condition, however, if you had high working memory capacity, you also had significantly higher odds of accurate responses (OR = 0.02, SE = 0.03, z = -2.370 p = 0.0180).

The model predicts that in the *non-impasse* control condition, the probability of a correct response for a participant with high vs. low working memory increases from (0.1 to 0.5%)—a negligible difference. In the *impasse* condition, however, the probability of a correct response increases from only 1% for participants with low working memory, to 42% for participants with high working memory. **These results are consistent with the intuition we develop from Figure 3.15. Participants with high working memory capacity were most able to take advantage of the impasse scaffold.** Raw data are visualized in Figure 3.15 [A], model predictions in Figure 3.15 [B] and parameter estimates and model specification are detailed in Appendix B.4.1.

STUDY 3C | ACCURACY

A Distribution | Question Accuracy



B Model | Probability of Correct Response

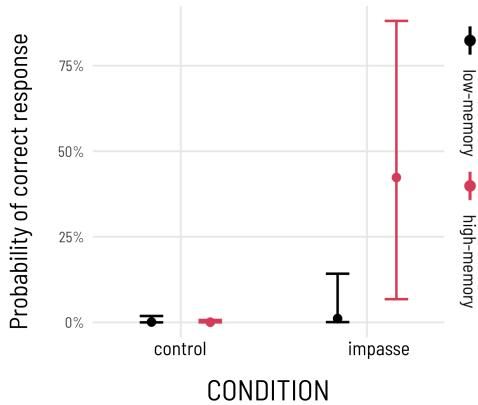


Figure 3.15. Study 3C – Accuracy. [A] A proportional bar chart of raw data shows that the proportion of correct responses increases most dramatically for individuals with high working memory capacity randomly assigned to the *impasse* condition. [B] The model predicts nearly the same probability of correct response regardless of WMC in the non-*impasse* control condition, and a significantly higher probability for the *impasse* condition.

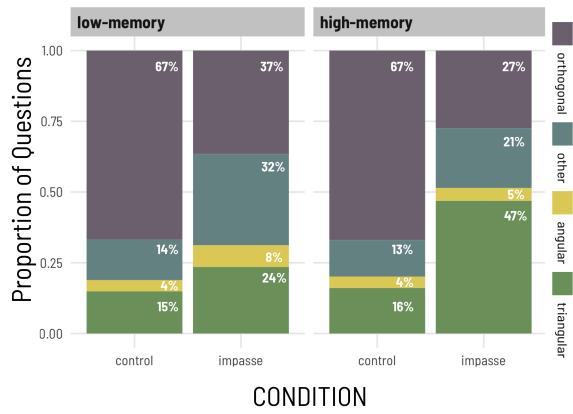
3.6.2.3 Interpretation: Does working memory affect graph interpretation?

To quantify the effect of working memory capacity on INTERPRETATION, we fit a Bayesian mixed multinomial regression model with random intercepts for subjects and questions. A Bayes Factor model comparison (against a random intercepts only model) indicates extreme evidence for the final model including fixed effects of **implicit scaffold**, WMC and their interaction term ($BF = 1.69e+13$). The model predicts similar probabilities for other, and angular interpretations across high vs. low working memory participants, indicating only a main effect of *impasse* scaffold increases probability of these transitional interpretations. It is only the (correct) *triangular* interpretation in which we have moderate evidence for a reliable interaction between WMC and **implicit scaffold** ($e^{\beta_{interaction}} = 15.73$, 95% CI [0.89, 249.91], $pd = 97.3\%$, $BF = 3.86$)

Consistent with the pattern of results for response accuracy, it is high working memory participants with higher probability of triangular responses, but only in the *impasse* condition. Raw data are visualized in Figure 3.16 [A], model predictions in Figure 3.16 [B] and parameter estimates and model specification are detailed in Appendix B.4.2.

STUDY 3C | INTERPRETATION

A Distribution | Question Interpretation



B Model Prediction | Probability of Interpretation

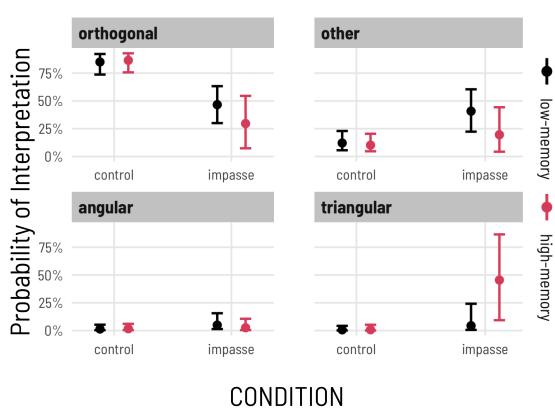


Figure 3.16. Study 3C – Interpretation. [A] A proportional bar chart shows that across both levels of the WMC factor, the proportion of correct *triangular* responses increases in the presence of the *impasse* structure. The model predicts a significantly higher probability of *triangular* responses, and similarly probabilities for *other*, and *angular* responses across *impasse* and explicit scaffolds.

3.6.3 Study 3C Discussion

In Study 3C we asked whether working memory capacity (WMC) could explain some of the individual differences in TM graph reading behaviour. Our results indicate that, although imposing an *impasse-structure* on the graph-reading problem significantly improves performance, this intervention might be most effective for individuals with *high* working memory capacity. Adding the *impasse* structure for individuals with *low* WMC (as measured by the OSPAN task) did not yield significantly more correct *triangular* responses. It did, however, yield more incorrect *transitional* interpretations (*angular*, *blank*, *other unidentifiable*); thus stepping away from cartesian and toward triangular coordinates.

Notably this pattern of results is different from that Lohse's (1997) empirical study finding that only individuals with low WMC whose performance improved with the addition of a better-designed graphical aid. Alternatively, we found that only individuals with high WMC were more accurate using the *impasse-structure*. We do not believe these findings are in conflict, however. Rather they highlight the difference between *using a graph* to perform an analytical

task (as in Lohse 1997), and *discovering* how a graph works, in order to perform an analytical task. Graph discovery is a prerequisite for accurate graph comprehension.

3.7 General Discussion

In this chapter, we've explored the graphical discovery component of graph comprehension though the theoretical lens of problem solving. We systematically explored the extent to which discovering the rules of a new graphical formalism proceeds like solving an insight problem, and sought to explain the behaviour of TM graph readers using the theoretical constructs of an information processing account of insight problem solving.

We find that imposing an impasse-structure on a TM graph reader significantly improves the odds of a correct *triangular* interpretation of the coordinate system, and further that these responses are most likely to occur in individuals with high working memory capacity. Thus we've found evidence pointing to one source of the substantial individual differences in TM graph reading performance. We also replicated our findings from Study 2 that providing explicit guidance (in the form of worked-example images) on how to read the graph also improves performance, and that in a static presentation medium where interaction is not available, adding an impasse-structure to an image scaffold will likely improve interpretation accuracy.

We also explored the variety of incorrect responses that readers offer to TM graph reading questions. In addition to the most common orthogonal (i.e. cartesian scatterplot) misinterpretation of the coordinate system, readers also provided responses that were blank or indicated uncertainty (e.g. selecting answer options consistent with both triangular and orthogonal interpretations), and some that were *angular*, indicating they recognized the importance of the diagonal grid and thought the orthogonal responses were inappropriate.

What do these alternative non-triangular and non-orthogonal responses tell us about the interpretive processing of the Triangular Model of Interval Relations? For participants in the impasse condition that offer *angular* responses, we argue it is likely they have reached an

impasse, accepted that an orthogonal interpretation is not appropriate, but failed to restructure their representation of the spatial relations between marks on the page *enough* or in the *correct way* to reach a complete triangular solution. Alternatively, blank responses likely indicate a participant encounters an impasse, accepts there is not an orthogonal solution, but is unable to conceive of an alternative, and thus leaves the answer blank. Other *unknown* type responses that do not correspond to any defined interpretation could also indicate this state of uncertainty, but perhaps these individuals did not think it appropriate to leave a question unanswered, and they select some (random) set of points to move forward to the next question. Alternatively these other responses could indicate some non-random though idiosyncratic interpretation we have not identified. A more certain account of these interpretations may be possible through interview or think-aloud protocols.

It is also important to note that *angular*, *blank* and *other* responses are also produced by participants in the control condition, though these responses are significantly less likely. This tells us that it is possible to experience a state of impasse even when an orthogonal answer is available. It is possible these individuals reach an impasse when they notice there are no orthogonal gridlines to traverse, and are perhaps less willing to violate the graph-reading norms we documented in Study 1 that are required to produce an orthogonal interpretation by mentally ‘superimposing’ a Linear Model atop the Triangular Model. Although there is much left to be discovered with respect to these alternative responses, they are significant in that they are *not* orthogonal. Any non-orthogonal but nonetheless incorrect response represents a positive step toward a more correct interpretation of the coordinate system.

Our prior work indicates that many TM graph readers apply procedures appropriate for other kinds of graphs (specifically, cartesian scatterplots) despite salient cues that these procedures may be inappropriate for the TM graph. In this chapter, we’ve explored how treating what is otherwise an *analytical problem*: applying the rules for extracting data from a graph; as a type of *creative* problem: discovering new rules; we’ve found an alternative, effective method for improving discovery of the TM graph’s interval coordinate system. Specifically, by anticipating

readers' cartesian misconception, and posing a question that deliberately leads to an impasse state, we've shown that more readers reach the correct graph interpretation. Accounts of insight problem solving would explain this result as a deliberate evocation of an impasse state leading to a restructuring of the reader's mental representation of the problem (i.e. the rules of the graphical formalism) allowing them to arrive at a correct solution.

Acknowledgements

This chapter, in part, includes portions of material as it appears in: Fox, Hollan, and Walker, 2019. *When Graph Comprehension Is An Insight Problem*. In Proceedings of the Annual Conference of the Cognitive Science Society. Additional material appearing in this chapter is being prepared for publication. The dissertation author was the primary investigator and author of these publications.

Chapter 4

Explorations of The Graphical Framework

Might readers avoid a mistaken interpretation altogether by instantiating a different *graph schema*? Our comprehension of novel representations is guided by our expectations based on prior knowledge of the conventions of graphical forms for a particular domain and presentation modality. In the case of the TM graph, our results suggest that expectations for the function of the coordinate system interfere with readers' ability to follow graphical cues provided by the graph's diagonal gridlines. One way to explore this influence is by applying the construct of the *graph schema*. Kosslyn (1989) posits a hierarchically organized graph schema is instantiated when reading a graph. But what schema is instantiated for a novel representation? Pinker (1990) speculates that upon encountering a novel graph, a reader will instantiate a *general graph schema* likely based on a combination of the graph's coordinate system and most salient graphical forms. For the TM graph, the tendency is to misinterpret the coordinate system as cartesian, mentally projecting (or physically tracing) non-existent orthogonal gridlines from the x-axis. In this way, we can describe misinterpretation of the TM graph as a failure to instantiate the appropriate graph schema, rather using an incorrect schema to read the graph. In this chapter, we explore what visuospatial features of the graph (gridlines, marks, gestalt shape, and orientation) might lead the reader to instantiate (or construct) an alternative graph schema, supporting discovery of the graph's novel coordinate system.

4.1 The Graphical Framework

Theory derived from empirical research in graph comprehension is consistent in the claim that successfully extracting information from a graph requires an integration of perceptual (i.e. ‘bottom-up’) and conceptual (i.e. ‘top-down’) cognitive processing (Carpenter and Shah, 1998; Hegarty, 2011; G. Lohse, 1993; Peebles and Cheng, 2002; Shah and Freedman, 2011)¹. That is, constructing meaning with a graphical formalism is much more than just *seeing*, it is an active, interpretive process that unfolds between the reader and physical artifact. To account for the role of prior knowledge in this interpretation (specifically prior knowledge of graphs) most contributions appeal to an all-important but vaguely-defined construct: the **graph schema**.

The first and most complete elaboration of the graph schema is given in Pinker’s (1990) *Theory of Graph Comprehension*². Pinker motivates his theory by addressing the proverbial elephant in the room. Despite the growing evidence that graphical displays present information in a fashion that is easier to reason with:

“...it is hard to think of a theory or principle in contemporary cognitive science that explains why this should be so; why, for example, people should differ so strikingly from computers in regard to the optimal input format for quantitative information” (Pinker, 1990, pg. 73).

His aim was to offer such a theory by elaborating the “cognitive operations” executed when reading a graph. This approach presupposes a representational account of cognition, with information processing proceeding via the propagation of representations in mind, the relative expressivity of propositions and language, but also a pragmatism—arguing that even though our experience with graphics may seem special upon introspection, it must do so by exploiting general purpose cognitive mechanisms.

¹At this (process) or task-analytic level of theories and models the distinction between perceptual and conceptual (all within the larger framework of cognition) is deliberately fuzzy. Rather, these terms as a shorthand to distinguish between what sources of information influence different theorized component processes. A change to the visuospatial design of a graph, for example would be expected to more strongly influence the bottom-up (stimulus driven) processes, while change to the expertise of the graph reader would be expected to more strongly influence the top-down (knowledge driven) processes.

²previously circulated as an MIT working paper as early as 1982

Pinker's theory rests upon a simple (task-analytic) observation he reasons from experience and appeal to Bertin (1967): that in reading a graph one must do two things. First, mentally represent the objects in the graph in *only* a particular way, and second, decide which aspects of the graph stand for which things the graph is trying to communicate. This parallels Palmer's (1978) characterization of 'representing' and 'represented worlds', and — examined critically—one can argue that these are not unique to graphing nor visual communication, but rather necessary conditions for any successful communicative exchange regardless of encoding or modality. One must determine which aspects of a signal represent the things we think the communicator wishes to exchange, and then decide upon a best interpretation of those aspects. Pinker maps these requirements into two distinct mental representations: (1) the **visual description**: encoding the depicted marks in terms of their spatial dimensions, and (2) the **graph schema**: how the spatial dimensions are mapped onto the measurement scales. Pinker further distinguishes the 'task' of graph reading in terms of: (1) a conceptual question: the particular sort of information a reader wishes to extract, and (2) a conceptual message: the actual information the reader takes away.

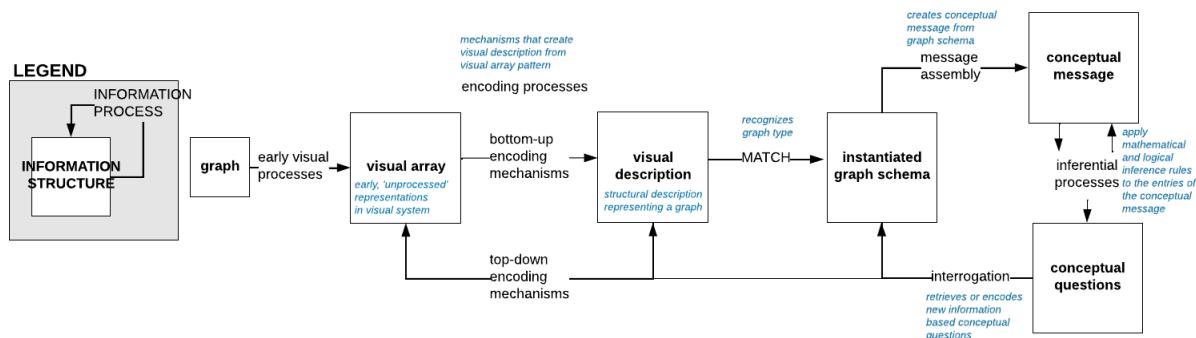


Figure 4.1. Summary of Pinker's Information Flow in Graph Comprehension. Adapted from Pinker, 1990 Figures 4.14 and 4.19, with colour annotations added by Amy Fox.

For Pinker, the graph schema fulfills a critical role in the larger process of graph comprehension, linking the representation constructed as a result of perceptual processing, with prior knowledge stored in long term memory. In his information processing account, (summarized in Figure 4.1) information enters the system as a visual array: a pattern of intensities on the retinas. In order to be useful, it needs to be transformed into symbolic representations —visual

descriptions —that can interface with representations in memory. Pinker steadfastly asserts that the visual description is symbolic in nature, appealing to Ullman (1984) and contemporaries to justify the translational mechanisms required. In this way, the visual description serves as a sort of intermediary representation, a symbolic schematic of the visual scene that can be operated upon in the context of reasoning. He proposes a subject-predicate structure where objects in a scene are parsed and represented symbolically via properties and relations. The visual description is constrained by a number of principles, grounded in “the totality of our knowledge of perception” (1990, pg. 78).

After the visual description is constructed for the scene, one still needs a way to interrogate it in relation to the conceptual question (information to be extracted). To fulfill this function, Pinker introduces the construct of the graph schema. He uses the term schema in a fashion consistent with contemporaries studying knowledge representation (Minsky, 1974; Norman and Rumelhart, 1975; Schank and Abelson, 1977): as a knowledge structure (in memory) containing a set of parameters and relations to be filled at a later time. Thus, the schema specifies both the properties that must be true of any objects of its class (i.e. via the relation between parameters) as well as what properties may vary from object to object (ie. the values of the parameters). The graph schema constrains the type and structure of information that can be instantiated, exerting a “top-down” influence on the interpretation of incoming “bottom-up” information.

4.1.0.1 Empirical Investigations of the Graph Schema

The first study to empirically examine the construct of the graph schema came from Ratwani & Trafton (2008) Using a mixing-costs paradigm, they compared the response times to first-order graph reading questions using horizontal & vertical bar charts, line, pie, and doughnut graphs. When participants were faster to answer questions in experimental blocks that contained graphs using the same coordinate system, the researchers concluded that the data supported an **invariant structure** view of the graph schema. According to this view, certain general characteristics (most notably the coordinate system) are shared across a number of graph types

that thus rely on a shared schema (Peebles and Cheng, 2003; Ratwani and Trafton, 2008). This is contrasted with a **perceptual feature view**, in which it is the surface features of the graph —the *graphical pattern* (e.g. lines, dots, bars, pies, etc.) that determines the graph schema (G. Lohse, 1993; G. Lohse, 1997). In the perceptual feature view, each type of graph has its own schema. Bar charts, line graphs, and pie charts would invoke different graph schemata, because their surface features (bars, vs. lines vs. circles) are different. Conversely, the invariant structure view would predict the instantiation of the same graph schema, because they rely on the same (cartesian) coordinate system. Ratwani & Trafton's mixing costs evidence that it takes longer for readers to read graphs when asked to switch between types using different coordinate systems is convincing insofar as it seems reasonable (and also consistent with both Pinker (1990) and Kosslyn's (1989) emphasis on the importance of the coordinate system). However, the evidence is less convincing as a refutation of the perceptual structure view. The graph schema is widely acknowledged as a hierarchical knowledge structure (i.e. Pinker, 1990; Ratwani, Trafton, and Boehm-Davis, 2008; Shah and Freedman, 2011), and it seems reasonable to expect that while highest levels of the hierarchy might be defined by the arrangement of marks in space (i.e. the coordinate system), lower levels are then likely to be defined by salient differences in perceptual features (i.e. lines vs. bars, or pies vs. donuts). That is to say, the two views do not seem entirely inconsistent.

4.1.0.2 The General Graph Schema

What schema is instantiated for a novel representation? Unfortunately neither the invariant nor perceptual structure views offer unique predictions as to what happens when the underlying coordinate system of a graph is not in reader's existing catalogue of schemata. Pinker argues that in the context of a particular situation, a specific graph schema is instantiated via a MATCH process, by which the visual description for a stimulus is compared with the the reader's available graph schemata (in long term memory), and the most appropriate schema is selected. It is this instantiated schema that that is then used to parse and extract information (conceptual

messages) from the graph, given the reader’s goals (conceptual questions). Pinker argues that upon encountering a novel graph, a reader will instantiate a “general” graph schema, likely based on a combination of the graph’s coordinate system and most predominate graphical forms (e.g. points, lines, bars, etc.) The exact mechanism of construction for this general schema is unknown, but Pinker suggests it may be related to the cognitive processes that represent abstract concepts like space and the movement of objects within it. It is likely, therefore, that any deep understanding of the genesis of a graph schema or structure of a generalized graph schema will implicate constructs from research on abstract thought, such as Conceptual Metaphor (Lakoff and Johnson, 1980) and Conceptual Integration Theory (Fauconnier and Turner, 2002).

This raises the important question: what graphical forms *indicate* to the reader the nature of the coordinate system? Conventional wisdom would suggest it is the axes: the marks which serve to orient the reader to the relationship between all the other marks. In Studies 4A-4C we explore whether changing the axes, and making even minor design changes to the other non-axis marks can lead more readers to discover the TM coordinate system.

4.2 Research Goals

In Studies 4A-D we systematically explore how visuospatial design features affect discovery of the coordinate system of a novel graph. In Study 4A we examine the role of **gridlines**, often considered a discretionary design feature, and ask if changing the extent of gridlines might lead readers to a different interpretation of the coordinate system. In Study 4B we explore the role of **marks**, features not considered a part of the graphical framework, and test whether altering the marks to reinforce the spatial relationships in the graph might improve discovery. In Study 4C we examine the gestalt **shape** of the graph space, evaluating whether changing the position of the y-axis from orthogonal (i.e. a square shape) to diagonal (i.e. a triangular shape) affects interpretation. Finally, in Study 4D we explore whether rotating the **orientation** of the entire graph in space better supports discovery. Our goal in each study is to

determine what features of the design might lead readers to instantiate a more appropriate graph schema, leading to the discovery of the novel graphical formalism.

4.3 Experiment 4A: Gridlines

Gridlines are falling out of fashion. In an age of graphical minimialism in visualization design, student students are encouraged to minimize or remove gridlines altogether. Famed popularizer of graphical displays Edward Tufte is often cited as characterizing gridlines as "chartjunk". His actual recommendation is more nuanced, and one with which this author agrees:

"One of the more sedate graphical elements, the grid should usually be muted or completely suppressed so that its presence is only implicit –lest it compete with the data. . . . Dark grids are chartjunk." Tufte, 1983, pg. 112.

But in addition to aiding graph readers in comparing the relative position of marks or reading specific values, I argue that gridlines serve an additional purpose: as reinforcing signals of the coordinate system. Although the primary signals of coordinate system are the *axes*, by guiding the reader's traversal of the graph space, the gridlines act as a sort of 'coordinate system training wheels'. This is certainly the case with the TM graph. In fact, in absence of the diagonal gridlines, with an orthogonal orientation of the y-axis, there would be no signal to the reader that the coordinate system is not cartesian, at all.

The design space (dimensions along which the feature might be altered) for gridlines is immense. For the TM graph in particular, we have decisions to make with respect to both the presence and density of both the x-axis diagonal gridlines and y-axis gridlines. In Study 4A we explore whether altering the design of the gridlines to make the graph appear *less* like a cartesian scatterplot might improve discovery of its coordinate system. We explore two alterations in particular (Figure 4.2).

In the *sparse horizontal* design, we remove the y-axis gridlines from areas of the graph in which data points cannot appear, because they are beyond the earliest start time or after the latest end time defined by the range of the x-axis. We refer to this as 'invalid interval space'. We

reason that by removing the gridlines from this area, we draw visual attention to the diagonal gridlines and make the triangular area of the graph more salient. In this way, the graph should appear less like a cartesian scatterplot and ideally facilitate the instantiation of a more general graph schema.

In the *full-diagonal* design we take the opposite approach. Rather than removing the horizontal gridlines from invalid interval space, we extend the diagonal gridlines *into* that space. This is a dense design, which Tufte might term as having a high 'data to ink' ratio, or even consisting of chartjunk. But we reason that by extending the diagonals into this space and creating a full square of diagonal lines: (1) the graph appears *very* different than a cartesian scatterplot, and (2) the grid becomes so dense it might interfere with mental projection of the orthogonal intersections required for computation of intervals under the linear-model interpretation of the graph.

The design of a grid is entirely discretionary, and so for the control condition we use the term *regular grid* only in reference to the fact that the design of the gridlines (diagonals on the hour, full-width horizontals on the hour) are the same as those used in each of our prior studies (1, 2, 3A-3C).

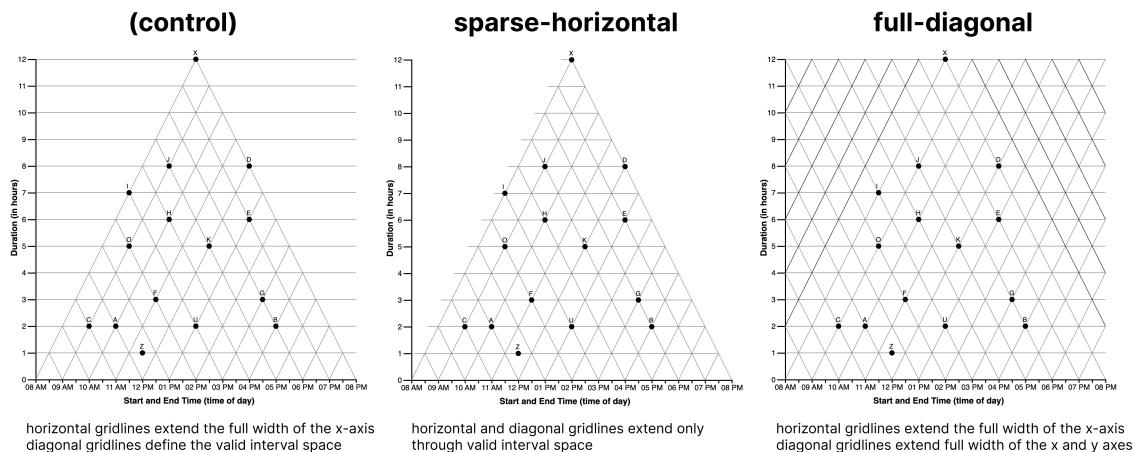


Figure 4.2. Study 4A – Grid Design Conditions. (at left) control condition; (at centre) sparse horizontal grid condition; (at right) full-diagonal grid condition.

Specifically, we hypothesize that:

- (H1) Removing the sections of the horizontal gridlines that cross ‘invalid interval space’ will emphasize the triangular shape, draw attention to the diagonal gridlines, and thus improve interpretation accuracy. Participants in the *sparse-horizontal* condition will have a higher probability of a correct response than participants in the *regular-grid* (control) condition.
- (H1) Extending the diagonal gridlines through ‘invalid interval space’ will make the graph appear less like a cartesian scatterplot, and thus improve interpretation accuracy. Participants in *full-diagonal* condition will have a higher probability of a correct response than participants in the *regular-grid* (control) condition.

4.3.1 Methods

4.3.1.1 Participants

We recruited 475 undergraduate students at UC San Diego to participate (online, asynchronously) in exchange for course credit. Twenty-two participants were excluded for failing attention-check questions or leaving the browser window during the study, yielding 453 participants for analysis (gender: 27 % male, 70 % female, 2 % other; age: 18 - 36 years).

4.3.1.2 Design

The experiment employed a multilevel design structure with 1 fixed, and 2 random factors:

(F1) **grid-design** (between-subjects) @ (c = 3) levels : *regular* [control], *sparse-horizontal*, *full-diagonal*

(R1) **question** (within-subjects) @ (q = 13) levels

(R2) **participant** @ (n = 453) levels

Participants were nested within grid-design condition, and questions were fully crossed with condition. Thus, each participant was randomly assigned to one grid-design, in which they completed all the questions (with the TM graph).

4.3.1.3 Materials

Interval Graph Comprehension Task

The Interval Graph Comprehension Task is similar to the task utilized in Study 3A described in section 3.3.1.3, differing only with respect to the block structure and implementation of the experimental manipulation.

The task begins by situating participants in a problem solving scenario where they are to assume the role of factory manager responsible for scheduling employee work shifts. They are instructed to complete the task by using a graph of the schedule to answer questions about the timing of shifts. A shift is defined as an interval of time with a discrete start and end time (on the hour). Thus the datapoints on the TM graph correspond to shifts, and shifts are identified with letters (i.e. A, B, C, etc). We chose to situate the task in the context of a scenario to give participants a familiar conceptual anchor for the interpretation of datapoints. Rather than the abstraction of ‘an interval of time’ each datapoint refers to a shift in an employee’s schedule.

The task proceeds with one experimental block, consisting of 15 ordered items. Participants complete the items in sequence, without feedback, and do not have the ability to skip ahead, nor return to previous items. In each item, the participant is presented with a TM graph, a question, and a grid of response options (see Figure 4.3). The same questions, scoring strategy, accuracy and interpretation measures were derived as described in Study 3. The design of the TM graph (shape, scale, axes, labels, size and relative screen position) *differs* by experimental condition.

Across all experimental conditions, participants complete the same sequence questions, with a graph visualizing the same dataset. Note that is different from Studies 3A-3C, where the application of the *impasse-structure* scaffold was accomplished by visualizing a different dataset, and only available during the first 5 questions. In Studies 4A-4D the dataset is *always* a non-impasse structure, in order to isolate any effects of manipulating graph design features. The experimental manipulation of **grid design** is accomplished by rendering a different set of gridlines for the graph. The three grids used in Study 4A are shown in Figure 4.2.

4.3.1.4 Procedure

Participants completed the study asynchronously over the internet, by accessing our custom web-application via the Chrome web-browser, with a keyboard and external mouse or trackpad (i.e. no mobile or touchscreen devices were permitted)³. After agreeing to an IRB-approved informed consent, participants were randomly assigned to an **grid design** condition and presented with task instructions. They then completed the Interval Graph Comprehension Task. Upon completion, participants were presented with a series of questions about their effort

³The stimulus web application renders the graph as 700x700 pixels. It detects browser window size and forces the browser into full screen mode. If the screen size is below the minimal threshold, participants are prevented from starting the task.

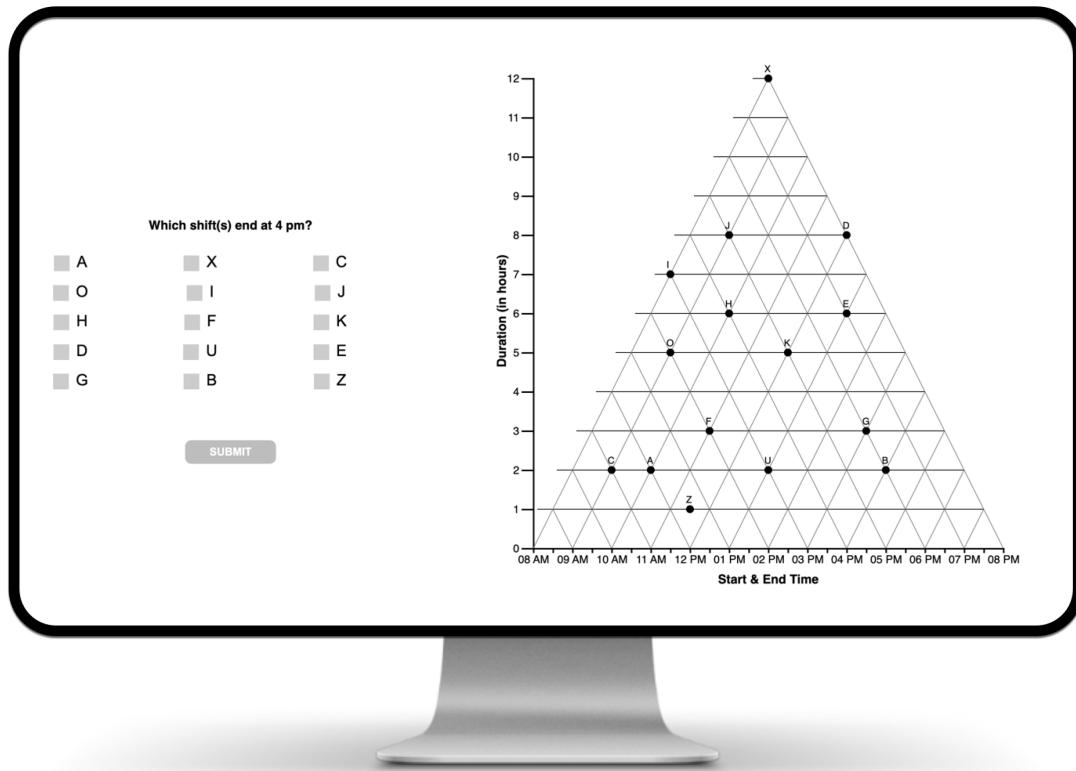


Figure 4.3. Study 4A – Layout of Graph Interpretation Task. See above the layout of the Interval Graph Interpretation Task (shown is Question #4; *sparse-horizontal* condition). Each item consists of a TM graph (at right), accompanied by a question, and grid of response options. Participants were instructed to check the boxes corresponding to all of the point in the graph that answered the question, and click the SUBMIT button to proceed to the next question.

and enjoyment of the task, followed by a demographic questionnaire, and final debriefing text.

4.3.1.5 Analysis

Response Accuracy

To test hypotheses related to response ACCURACY we fit mixed logistic regression models (generalized linear mixed models (GLMM) with a logistic link function) in R using the `lme4` package (Bates, Mächler, Bolker, and Walker, 2015; R Core Team, 2022). Note that we choose to model these data at the question rather than participant (i.e. TOTAL SCORE) level because the structure of a mixed effects model allows us to differentiate between random variance introduced by individual participants and questions, versus the expected systematic variance of experimental condition. Further, the distribution of total accuracy score at the participant level was bimodal, and violated the assumptions of normally distributed residuals and homogeneity of variance required by OLS linear regression. For contrast coding categorical variables, the default treatment (dummy) coding scheme was used: on the response variable ACCURACY the level *0:incorrect* was defined as the reference category, and on the predictor variable **grid design** the level *regular* (control) was defined as the reference category. Thus exponentiated model intercept e^{b_0} refers to the baseline odds of a *correct* response in the *regular* (control) condition, while exponentiated model coefficient e^{b_1} refers to the odds-ratio (relative increase or decrease in odds) of a *correct* response for each level of the non-control experimental conditions. To determine a final model we first defined the maximal random effects structure theoretically justified by the study design (random intercepts for questions and subjects). We then fit a model with **grid-design** as predictor and used a likelihood ratio test to decide if adding the predictor resulted in a significantly better fit. Statistical significance of each predictor in the final model was determined via Wald Chi-Square tests, and all reported p-values are for non-directional tests with a decision threshold $\alpha = 0.05$.

4.3.2 Results

4.3.2.1 Total Score

To explore the effect of **gridlines** on TM graph reading performance, we start by describing the distribution of TOTAL SCORE. Across all conditions, TOTAL SCORE ranged from 0 to 100 with a mean of 15%. In Figure 4.4 we see that across all gridline conditions, participant level accuracy on the interval graph comprehension task is low (less than 50%). The full grid condition has the lowest mean score (10%) and the smallest variance.

STUDY 4A | Distribution of Total Score

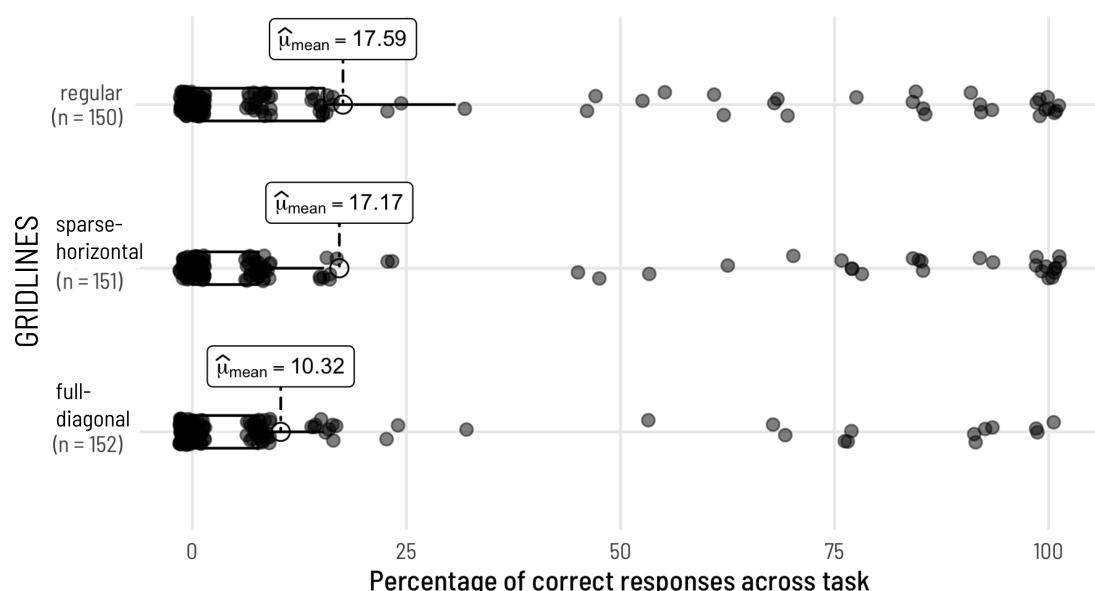


Figure 4.4. Study 4A – Distribution of Total Score. The mean score across all three grid designs is very low. Rather than increasing performance, the the *full diagonal grid* yielding the lowest scores overall.

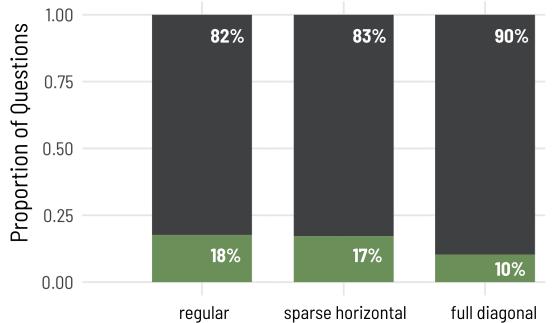
4.3.2.2 Question Accuracy

To explore the effect of gridlines on accuracy, we fit a mixed effects logistic regression model with random intercepts for subjects and questions, with gridline design as a fixed effect. A likelihood ratio test indicates that a model including these main effects does not significantly

more variance in the data than an intercepts-only baseline model ($\chi^2(3, 5) = 0.82, p = 0.66$).

STUDY 4A | ACCURACY

A Distribution | Question Accuracy



B Model | Probability of Correct Response

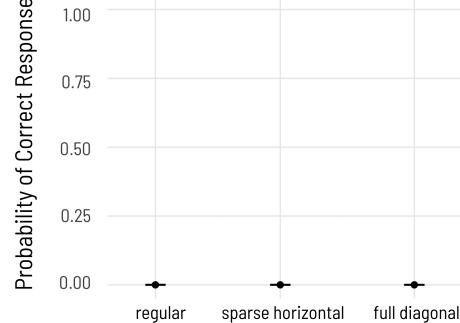


Figure 4.5. Study 4A – Accuracy. [A] A proportional bar chart of raw data shows that the proportion of correct responses increases is low across all **grid** designs. [B] The model predicts less than 1% probability of a correct response in each condition.

Counter to our (H1) hypothesis, altering the design of the horizontal gridlines to emphasize the triangular portion of the graph space does not significantly improve accuracy. Model coefficients indicate that relative to the full y-axis grid condition, removing gridlines outside the eligible data area does not increase the odds of a correct response ($e^{b_{1[sparse]}} = 1.06, SE = 0.75, p = 0.94$).

Counter to our (H2) hypothesis, altering the design of the diagonal gridlines to obscure the triangular shape but make the diagonal grid more prominent does not improve accuracy. Extending the diagonal grid through the entirety of the square graph space defined by the x and y axes does not significantly change the odds of a correct response ($e^{b_{1[grid]}} = 0.59, SE = 0.41, p = 0.43$). Model predictions are visualized in Figure 4.5 [B], while parameter estimates and model specification is detailed in Appendix C.1.1.

4.3.3 Study 4A Discussion

In Study 4A we found that neither of our gridline manipulations succeeded in improving interpretation of the TM graph. The full grid design, in fact, decreased accuracy, though not to a

degree which was beyond the range of sampling variability.

One interpretation of these results is that the gridlines are not, as we've argued, indicators of the coordinate system. But we believe the more likely explanation is that our design manipulation of was not strong enough or in the precise way to help readers discover the coordinate system. In our observational Study 1, we found video evidence for readers using the gridlines to traverse the graph space. In Study 3A, we found evidence via mouse cursor tracking that readers used the gridlines to traverse the graph space. Thus, although gridlines may help us traverse the graph area in the correct way, it seems that altering their design is not a silver bullet solution to improving coordinate system discovery.

4.4 Experiment 4B: Marks

Marks can be powerful indicators of semantic content, though in comparison to the broader range of graphic external representations including infographics and diagrams, in the design of graphs and charts, variance in marks is more limited. In statistical graphs in particular, mark shape is most commonly used as a redundant-encoding for colour in order to support accessibility. For example, differently shaped marks might be used to encode an additional variable in a bivariate scatterplot, turning it into a bivariate scatterplot with two or more groups of data, where the groups are double-encoded by both colour and shape.

In Study 4B we explore whether we can use the shape of a mark (indicating the intervals data points on the TM graph) to signal to the reader the relationship between the marks and the x-axis: the most novel component of the TM graph. We compare three design alternatives (Figure 4.6).

In the *arrow* design, we replace the default point marks with small arrows, oriented such that the tails are aligned with the diagonal gridlines, intended to draw attention from the point of the arrow, down toward the x-axis. We reason this might act as a visual scaffold leading the reader to trace down *both* gridlines intersecting a point, and thus discover the coordinate system.

In the *cross* design we replace the default point with small cross. This serves as a check on novelty. The cross design is less common than a point shaped mark, but does not reinforce the relationship between the data point and the x-axis.

The control condition consists of the point marks used in each of our prior studies prior studies (1, 2, 3A-3C, 4A).

Specifically, we hypothesize that:

- (H1) Altering the design of the mark indicating an interval so as to emphasize the relationship between the mark and the diagonal gridlines will improve interpretation accuracy. Participants in the *arrow* condition will have a higher probability of a correct response than participants in the *point* (control) condition.
- (H2) Altering the design of the mark in a way that is unconventional (i.e. less like a cartesian scatterplot) but that does not emphasize the relationship between the mark and x-axis will *not* improve interpretation accuracy. Participants in the *cross* condition will not have a higher probability of a correct response than participants in the *point* (control) condition.

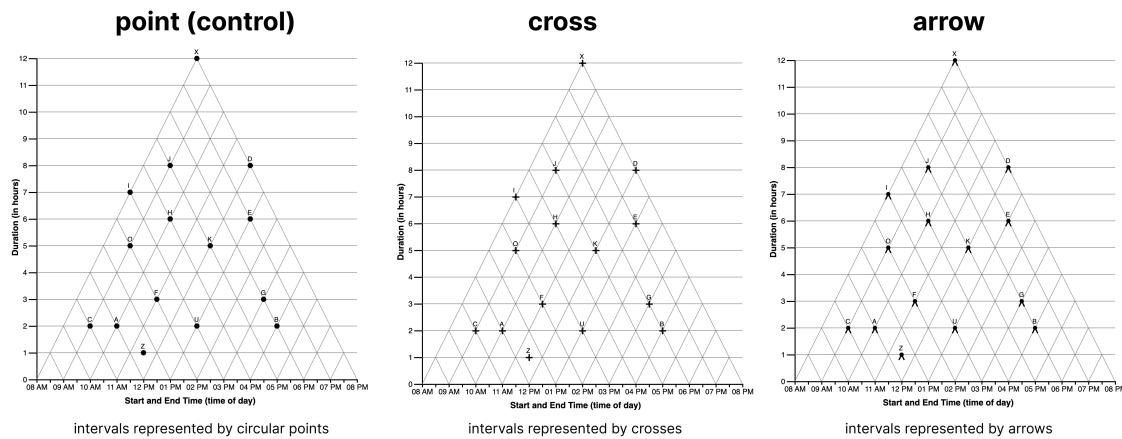


Figure 4.6. Study 4B – Mark Design Conditions. (at left) point [control] condition; (at centre) cross condition; (at right) arrow condition.

4.4.1 Methods

4.4.1.1 Participants

We recruited 330 undergraduate students at UC San Diego to participate (online, asynchronously) in exchange for course credit. Twenty-nine participants were excluded for failing attention-check questions or leaving the browser window during the study, yielding 301 participants for analysis (gender: 36 % male, 63 % female, 2 % other; age: 18 - 30 years).

4.4.1.2 Design

The experiment employed a multilevel design structure with 1 fixed, and 2 random factors:

(F1) **mark-design** (between-subjects) @ (c = 3) levels : *point* [control], *cross*, *arrow*

(R1) **question** (within-subjects) @ (q = 13) levels

(R2) **participant** @ (n = 301) levels

Participants were nested within mark-design condition, and questions were fully crossed with condition. Thus, each participant was randomly assigned to one grid-design, in which they completed all the questions (with the TM graph).

4.4.1.3 Materials & Procedure

The same task, scoring, measures and procedure were used as described in Study 4A.

4.4.1.4 Analysis

Response Accuracy

To test hypotheses related to response ACCURACY we fit mixed logistic regression models in R using the lme4 package. For contrast coding, the default treatment (dummy) coding scheme was used with the following reference categories:

- response variable ACCURACY : level 0:*incorrect* as reference

- predictor factor **mark design**: level *point (control)* as reference

To determine a final model we first defined the maximal random effects structure theoretically justified by the study design (random intercepts for questions and subjects). We then fit a predictor model with **mark design** and used a likelihood ratio test to determine if this model was superior to a simpler model including random effects only. Statistical significance of each predictor in the final model was determined via Wald Chi-Square tests, and all reported p-values are for non-directional tests with a decision threshold $\alpha = 0.05$.

4.4.2 Results

4.4.2.1 Total Score

To explore the effect of **marks** on TM graph reading performance, we start by describing the distribution of TOTAL SCORE. Across all conditions, TOTAL SCORE ranged from 0 to 100 with a mean of 21%. In Figure 4.7 we see that participant level accuracy on the interval graph comprehension task is low (less than 50%) consistent with prior studies, but that the average score (and variance in the distribution) are slightly higher for the *cross* and *arrow* conditions.

4.4.2.2 Question Accuracy

To explore the effect of **marks** on ACCURACY, we fit a mixed effects logistic regression model with random intercepts for subjects and questions, with mark design as a fixed effect. A likelihood ratio test indicates that a model including this main effect is a significantly better fit for the data than an intercepts-only baseline model ($\chi^2(3,5) = 15.09, p < 0.001$). A Wald Chi-Square tests confirms a main effect of mark design ($\chi^2(2) = 12.7, p < 0.001$).

Consistent with our (H1) hypothesis, altering the design of the marks to emphasize the relationship between the point and the diagonal gridlines did significantly improve accuracy. Model coefficients indicate that relative to the point mark condition, the arrow mark increases the odds of a correct response ($e^{b_{1[arrow]}} = 31.95, SE = 3.54, p < 0.001$).

Consistent with our (H2) hypothesis, altering the design of the mark in a way that

STUDY 4B | Distribution of Total Score

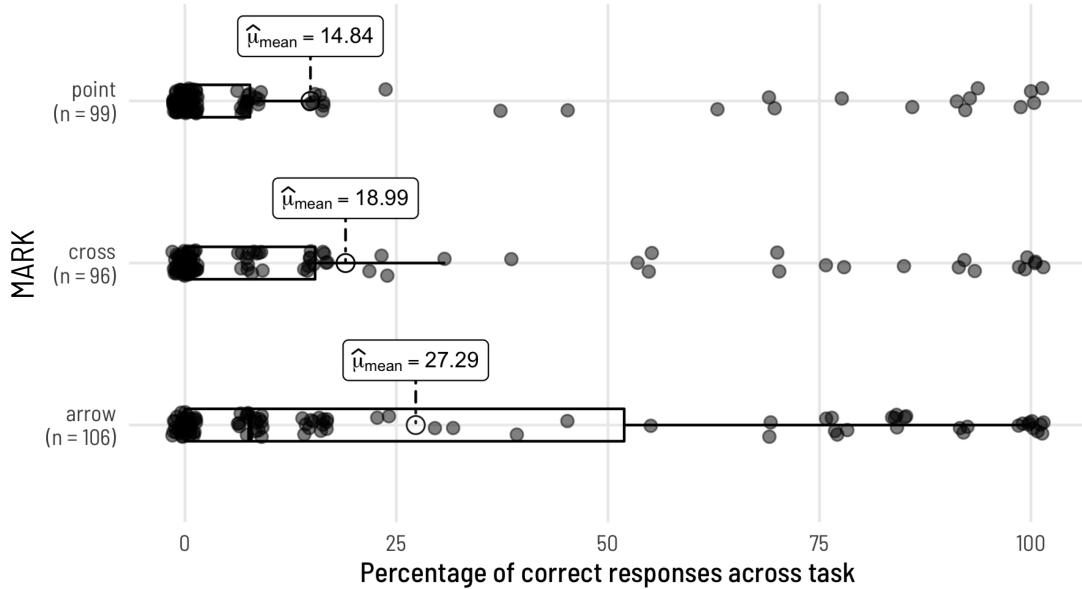


Figure 4.7. Study 4B – Distribution of Total Score. Only the *arrow* condition yields a substantial increase in total score.

STUDY 4B | ACCURACY

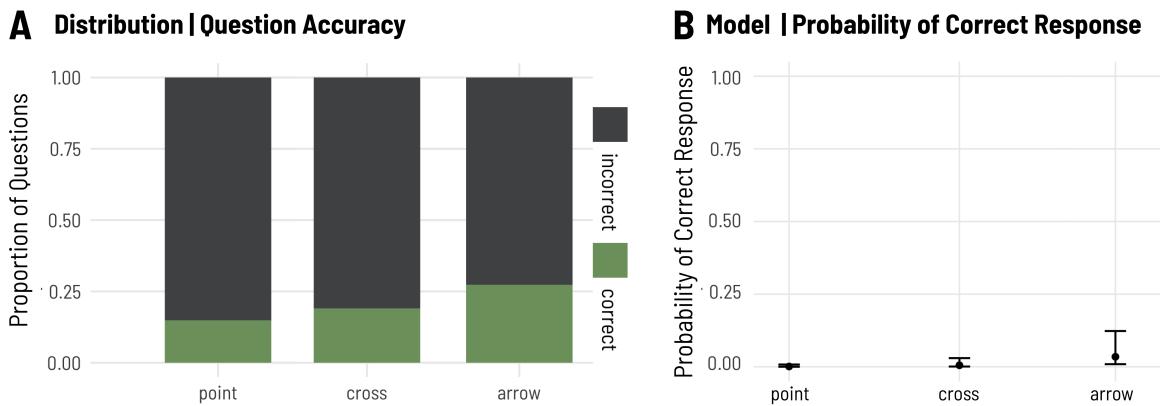


Figure 4.8. Study 4B – Accuracy. [A] A proportional bar chart of raw data shows that the proportion of correct responses increases is low across most **mark** designs, with a small increase in the *arrow* condition.

is unconventional, but that does not emphasize the relationship between a point and the gridlines does not improve accuracy. Model coefficients indicate that relative to the point mark condition, the cross mark did not change the odds of a correct response ($e^{b_1[\text{cross}]} = 4.71, SE =$

$4.48, p = 0.10$). Model predictions are visualized in Figure 4.8 [B], while parameter estimates and model specification is detailed in Appendix C.2.1.

4.4.3 Study 4B Discussion

Results from Study 4B support our hypotheses that in addition to supporting semantic content (such as reinforcing a categorical coding scheme) the design of a mark can be used as a signal to the functioning of a novel coordinate system. Specifically, changing from a point to a small arrow design for the marks on the TM graph lead to small increase in the probability of correct interpretation. We believe this small design change may function in a way similar to that of the image-based scaffolds from Study 2 and Study 3B, to a lesser extent. The arrow mark simply hints at the need to attend to the diagonal gridlines, while highlighting the intersection between a datapoint and the x-axis via the gridlines is a more explicit demonstration, and one that is hard to ignore.

4.5 Experiment 4C: Shape and Scale

The most powerful indicator of a coordinate system is its axes: the marks designed to orient the reader to how all other marks in the space are related. In the TM graph, these are a horizontal x-axis, and in the version documented by (Kulpa, 2000; Qiang, Delafontaine, Versichele, De Maeyer, and Van de Weghe, 2012; Van de Weghe, Docter, De Maeyer, Bechtold, and Ryckbosch, 2007) a vertical y-axis. As we noted in Study 4A, however, any space outside the bounding triangle of the TM graph defined by the starttime-endtime range of the x-axis, is not valid interval space. That is, no datapoints can appear in that space. And thus, there is no reason, strictly speaking, that the y-axis of the graph needs to be *orthogonal* to the x-axis. The y-axis could be oriented as the left-most (i.e. earliest) ascending diagonal gridline. One particularly clever participant in our Study 2 drawing task spontaneously produced this design improvement. We reason that this is highly likely to greatly improve graph discovery because perhaps the orthogonality between the x and y axis is the visual property responsible for triggering a the

instantiation of a cartesian graph schema. Removing the orthogonal y axis makes it the graph appear less like a cartesian scatterplot, and increases the salience of the diagonal gridlines .

A second discretionary design aspect of the TM graph is the angle at which the diagonal gridlines meet the x-axis. As (Van de Weghe, Docter, De Maeyer, Bechtold, and Ryckbosch, 2007) notes, they choose an angle to be consistent with Kulpa's formalization in (Kulpa, 1997, 2001), but that this angle is in fact arbitrary, and changing it does not change the nature of the coordinate system or relation between the intervals in space, it merely scales them. Scaling the angle however does do two things: (1) it changes the gestalt shape of the bounding triangle, and (2) changes the scale between the x and y axes. In studies 1-4B we used an isosceles shape to the bounding triangle that afforded a 1:1 scale between the distance between 1 hour on the x axis and 1 hour on the y axis. We wonder, however, if this scaling might be unconsciously processed by the perceptual system and lead readers toward an orthogonal interpretation, because the in our isosceles design, 1 hour altitude (i.e an orthogonal projection) of the triangle is equal to 1 hour on the x-axis. To change the scale of the x and y axes such the gestalt shape of the bounding triangle is equilateral means that it there is a 1:1 mapping between 1 hour of distance on the x axis and 1 hour of distance on the *diagonal gridline*: the traversal we want readers to make.

In Study 4C, we evaluate the effect of changing the shape of the graph (via orientating the y-axis) and scale, on interpretation of the TM graph. We explore the following combination of design factors (Figure 4.9).

Specifically, we hypothesize that:

(H1) Reducing the graph area to only ‘valid interval space’ by collapsing the y-axis to the first diagonal gridline (a triangular shape) will improve discovery. Participants in the *triangular* condition will have a higher probability of a correct response than participants in the *square* (control) condition.

(H2) Re-scaling the angle of the diagonal grid (and thus changing the type of triangular shape)

will not improve performance. Participants in the *equilateral* condition will not have a higher probability of a correct response than participants in the *isosceles* (control) condition.

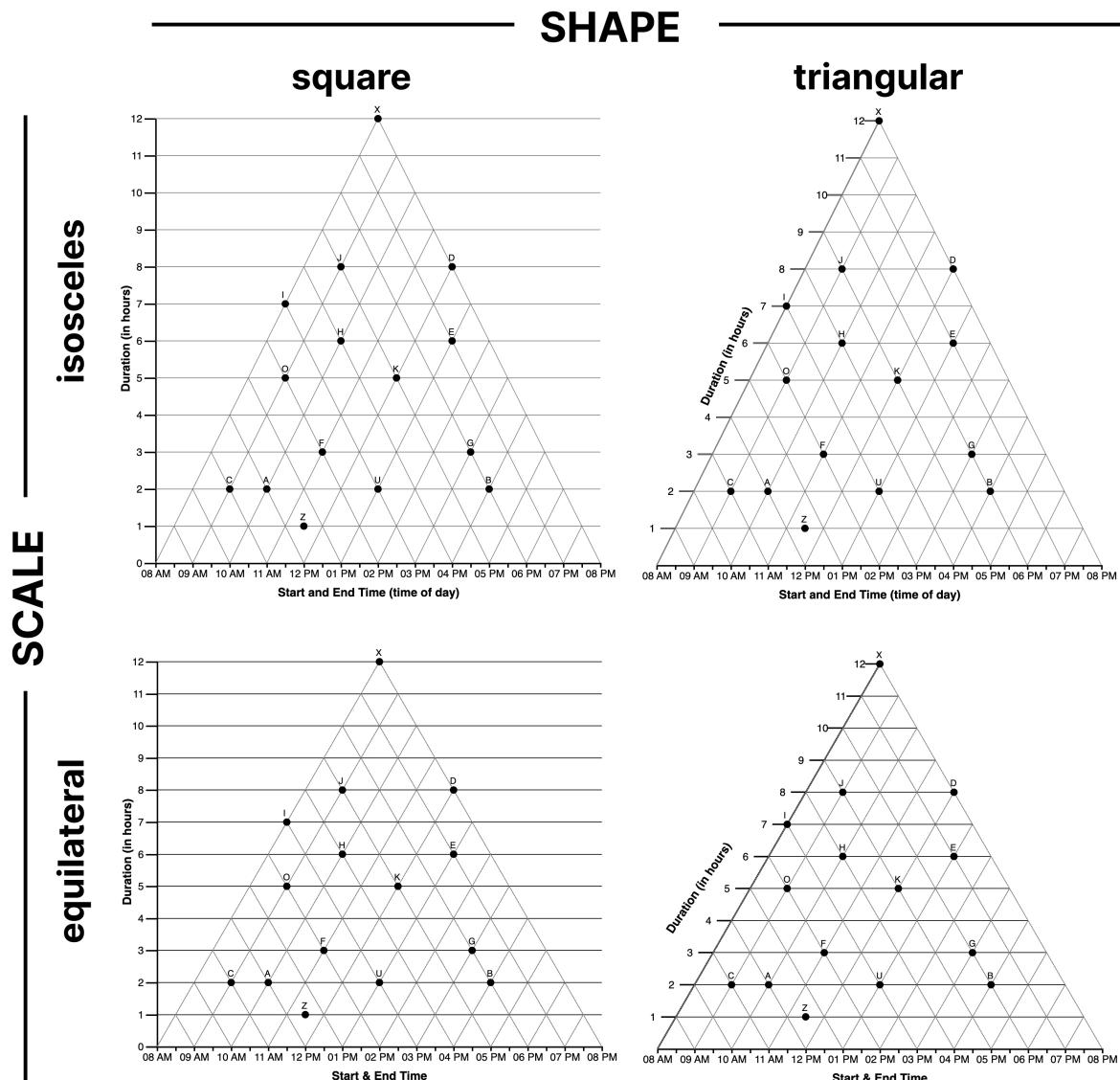


Figure 4.9. Study 4C – Shape & Scale Design Conditions. The top row shows the *isosceles* conditions, and the bottom row the *equilateral* across both the *square* shape (left column) and *triangular* shape (right column).

4.5.1 Methods

4.5.1.1 Participants

We recruited 244 adults located in the United States via the Prolific subject recruitment platform to participate in exchange for monetary compensation. Five participants were excluded for failing attention-check questions or for leaving the browser window during the study, yielding 239 participants for analysis (gender: 42 % male, 55 % female, 3 % other; age: 18 - 71 years).

4.5.1.2 Design

The experiment was defined by a multilevel factorial structure with 2 fixed and two random factors:

(F1) **shape** (between-subjects) @ (c = 2) levels: *square* [control], *triangular*

(F2) **scale** (between-subjects) @ (c = 2) levels : *isosceles* [control], *equilateral*

(R1) **question** (within-subjects) @ (q = 13) levels

(R2) **participant** @ (n = 239) levels

The two fixed factors were fully crossed, yielding four distinct conditions: *square |isosceles*, *square |equilateral*, *triangular |isosceles*, *triangular |equilateral*. Participants were nested within condition, and questions were fully crossed with condition. Thus, each participant was randomly assigned to one of the four (factorial) conditions, in which they completed all questions.

4.5.1.3 Materials & Procedure

The same task, scoring, measures and procedure were used as described in Study 4A.

4.5.1.4 Analysis

Response Accuracy

To test hypotheses related to response ACCURACY we fit mixed logistic regression models in R using the lme4 package. For contrast coding, the default treatment (dummy) coding scheme

was used with the following reference categories:

- response variable ACCURACY : level *0:incorrect* as reference
- predictor factor **shape**: level *square (control)* as reference
- predictor factor **scale**: level *isosceles (control)* as reference

To determine a final model we first defined the maximal random effects structure theoretically justified by the study design (random intercepts for questions and subjects). We then fit the most complex model indicated by the study design, including (main) fixed effects for **shape** and **scale** scaffold as well as their interaction term, and used a likelihood ratio test to determine if this model was superior to a simpler model including fixed main effect only. Statistical significance of each predictor in the superior model was determined via Wald Chi-Square tests, and all reported p-values are for non-directional tests with a decision threshold $\alpha = 0.05$.

4.5.2 Results

4.5.2.1 Total Score

To explore the effect of **shape** and **scale** on TM graph reading performance, we start by describing the distribution of TOTAL SCORE. Across all conditions, TOTAL SCORE ranged from 0 to 100 with a mean of 33%. In Figure 4.10 we see that participant level accuracy on the interval graph comprehension task is low (less than 50%) consistent with prior studies. We also see that scores were lowest for the **equilateral** scale conditions, across both graph shapes. For the **isosceles** scale however, variance was substantially greater for the *triangle* shaped graph, indicating that more participants discovered the coordinate system at least part way through the task.

4.5.2.2 Question Accuracy

To explore the effect of **shape** and **scale** on ACCURACY, we fit a mixed effects logistic regression model with random intercepts for subjects and questions, with **shape** and **scale** as

STUDY 4C | Distribution of Total Score

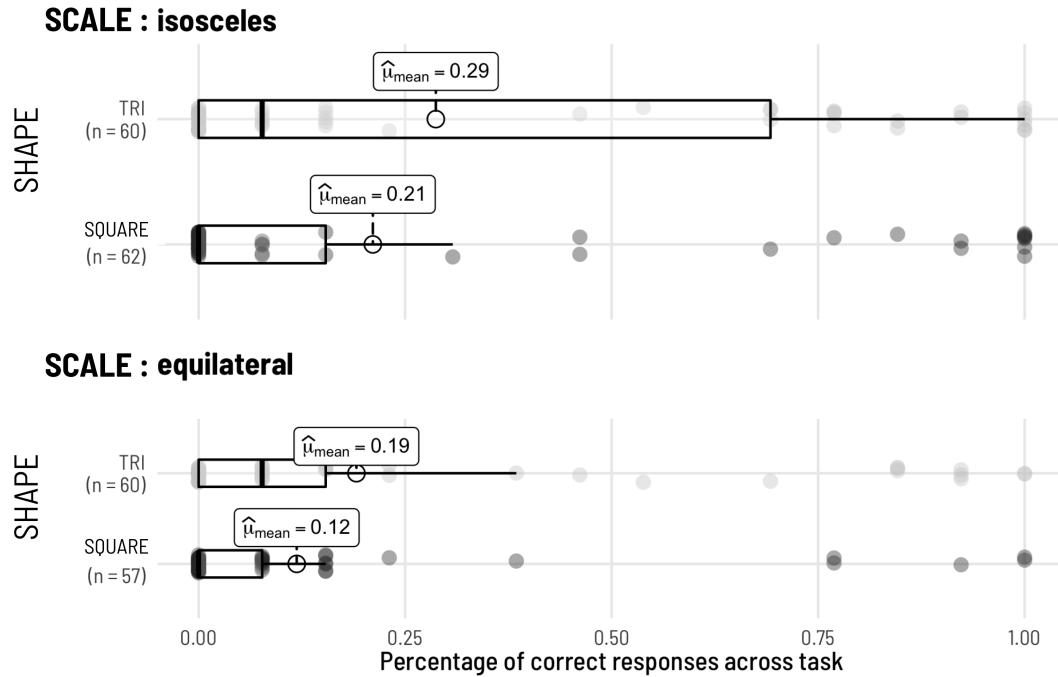


Figure 4.10. Study 4C – Distribution of Total Score. Re-orienting the y-axis to create a triangular rather than square graph shape results in a small but significant increase in accuracy, but re-scaling the triangles from isosceles to equilateral does not.

fixed effects. A likelihood ratio test indicates that a model including these main effects explains significantly more variance in the data than an intercepts-only baseline model ($\chi^2(3,5) = 3.41, p < 0.05$). We also fit a model including an interaction term between **shape** and **scale**, however a likelihood ratio test indicated that adding the interaction term did not improve model fit ($\chi^2(5,6) = 0.51, p = 0.47$) Therefore we chose the simple main effects model (with random intercepts) as the final model. The explanatory power of the entire model is substantial (*conditional R²* = 0.93) though the part related to the fixed effects **shape** and **scale** (*marginal R²*) explains only 2% of variance.

Wald Chi-Square tests revealed significant main effects for shape ($\chi^2(1) = 3.17, p < 0.05$, one-tailed), but no main effect for **scale** ($\chi^2(1) = 0.63, p = 0.22$, one-tailed). **Consistent with our (H1) hypothesis, a triangular y-axis improves accuracy relative to an orthogonal y-axis.**

STUDY 4C | ACCURACY

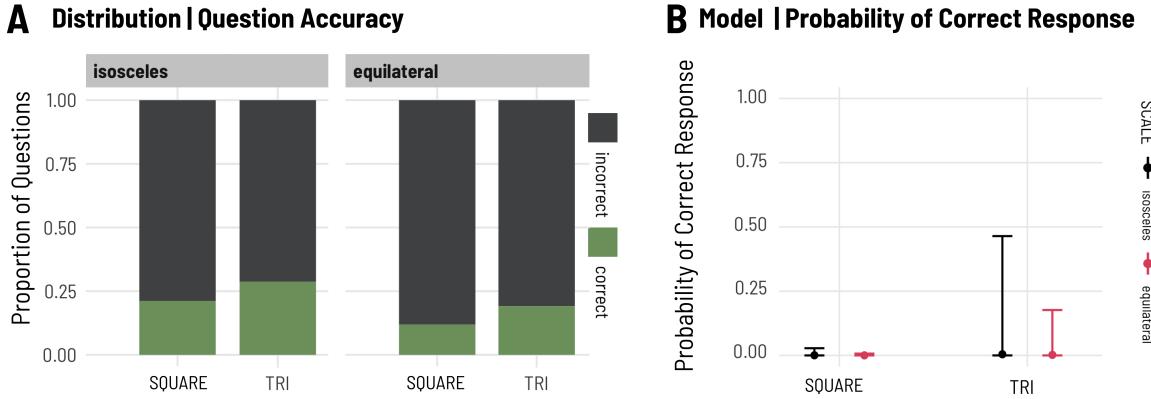


Figure 4.11. Study 4C – Accuracy. The model predicts a small but significant increase in accuracy for the triangular (vs) square shaped graph.

Model coefficients indicate that across both isosceles and equilateral scales, collapsing the y-axis from orthogonal to triangular increases the odds of a correct response by a factor of 8 ($e^{b_1[\text{triangular}]} = 7.77, SE = 8.95, p < 0.05$).

Consistent with our (H2) hypothesis, re-scaling the diagonals from isosceles to equilateral does not improve accuracy. Across both axis shapes, re-scaling the graph from isosceles to equilateral does not significantly change the odds of a correct response ($e^{b_1[\text{equilateral}]} = 0.47, SE = 0.44, p = 0.22$). Model predictions are visualized in Figure 4.11 [B], while parameter estimates and model specification is detailed in Appendix C.3.1.

4.5.3 Study 4C Discussion

In Study 3C we find evidence that changing the overall shape of the graph space from *square* to *triangular* has a small but significant effect on improving interpretation accuracy. From a design perspective, it is a more spatially-efficient and parsimonious choice, as it removes from the graph area space that cannot be used to represent intervals. Re-scaling the shape of the triangular grid from isosceles to equilateral, however, did not improve (and in fact for this sample, slightly decreased) accuracy, across both graph shapes. These results imply that, consistent with the definition of the graphical formalism, the angle of the gridlines (which defines the relative

scale of the x and y axes) does not communicate meaning to reader. Importantly, although the equilateral shape more faithfully represents the 1:1 relationship between 1 hour on the x and y axes as being the same distance using the diagonal grid, this is not information that the reader seems to make use of.

4.6 Experiment 4D: Orientation

Studies 1-4 offer substantial evidence of the orthogonal tracing behaviour we believe gives rise to the most common, incorrect interpretation of the TM coordinate system. Despite prominent cues (including explicit instructions), some readers persist in this mistaken interpretation. In Study 3A (the impasse hypothesis), we tried to stop readers from making this cartesian mistake by asking a question for which making the orthogonal projection from the x-axis would not yield an available answer. In Study 4D we similarly try to prevent this mistake, this time by making it *harder* to trace an orthogonal projection. Specifically, we believe that by rotating the graph in space by 45 degrees, we can make it harder for readers to mentally project a gridline onto the graph that does not exist. To isolate inhibitory affect on mental projection that rotation of the graph might exert, we compare this against a third condition in which we rotate the graph by 90 degrees.

We reason that, supported by the horizontal and vertical edges of the computer screen, it is (relatively) easy for readers to mentally project horizontal and vertical lines onto the graph, but more challenging to project diagonal lines, especially ones that are not parallel to the existing diagonal gridlines drawn on the graph. Thus, by rotating the graph, we expect it will be more difficult for readers to perform the incorrect operation, which will hopefully make them more likely to make use of the lines drawn on the graph, and arrive at a correct interpretation.

We explore the following combination of factors (Figure 4.12). Note that rotating the graphs requires design decisions regarding the relative orientation of axis labels, tick mark labels and data labels. Across graph shapes in the 45 degree rotation condition, we choose to keep both

x and y axis tick mark labels oriented in parallel with the gridlines to which they correspond. We reason this will minimize the likelihood that readers will simply *tilt their head* to read the graph. We kept the x and y axis labels, however, parallel to the x and y axes, as it is possible to read these labels at that angle without tilting one's head. In the 90 degree rotation condition we rendered the x axis (now oriented vertically in space) perpendicular to the the axis. This choice was made, again, to avoid encouraging the reader to shift their head to read the labels, and also to remain neutral as to whether a given intersection with the axis represents a start time (i.e. downward directed gridline) or end time (i.e. upward directed gridline). In all cases the data labels were rotated to be vertical with respect to the display screen (thus not requiring any head shifting to read them clearly).

Specifically, we hypothesize that:

- (H1) Rotating the both the *triangular* and *square* shaped graphs in space will improve interpretation of the TM graph. We predict that the *45 degree* rotation will be more effective than the *90 degree* rotation, which will nonetheless be superior to the *0 degree* (non-rotated control).
- (H2) Rotation will be effective across both graph *triangular* and *square* shapes.
- (H3) In replication of the effect of **shape** in Study 4C, participants will have a higher probability of accurate interpretation for the *triangular* rather than *square* shaped axes.

4.6.1 Methods

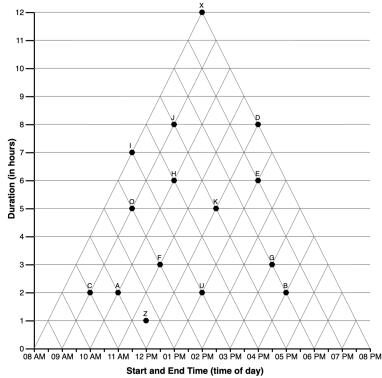
4.6.1.1 Participants

We recruited 412 adults located in the United States via the Prolific subject recruitment platform to participate in exchange for monetary compensation. Twenty one participants were excluded for failing attention-check questions or for leaving the browser window during the study, yielding 391 participants for analysis (gender: 42 % male, 55 % female, 3 % other; age: 18 - 79 years).

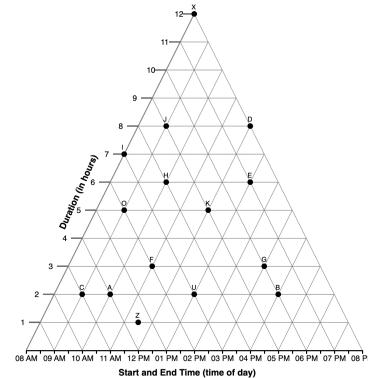
SHAPE

square

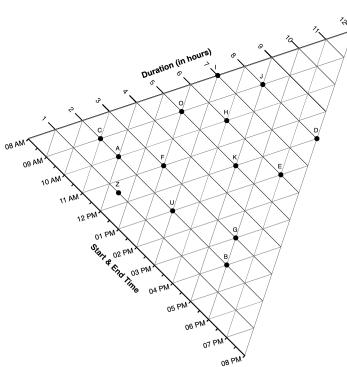
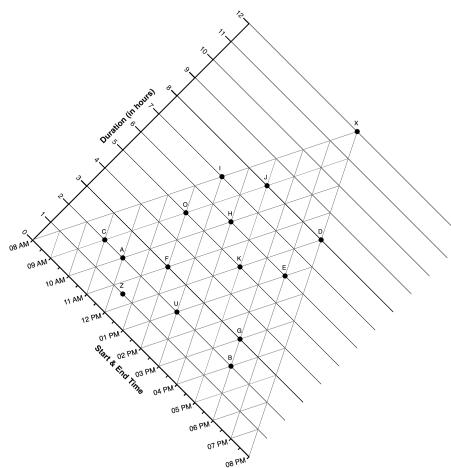
0 degree



triangular



ROTATION
45 degree



90 degree

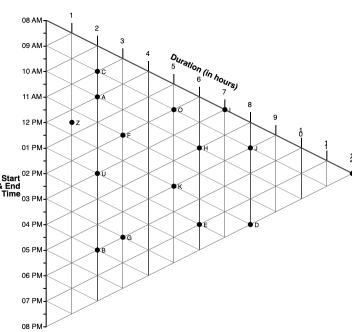
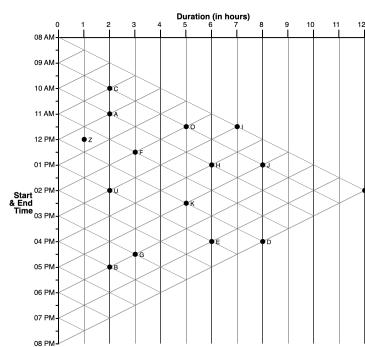


Figure 4.12. Study 4D – Shape & Rotation Design Conditions. The top row shows the *0 degree* (control) conditions, centre row the *45 degree* rotation and the bottom row the *90 degree* rotation, across both the *square* shape (left column) and *triangular* shape (right column).

4.6.1.2 Design

The experiment was defined by a multilevel factorial structure with 2 fixed and two random factors:

(F1) **shape** (between-subjects) @ (c = 2) levels: *square* [control], *triangular*

(F2) **rotation** (between-subjects) @ (c = 3) levels : *0 degrees* [control], *45 degrees*, *90 degrees*

(R1) **question** (within-subjects) @ (q = 13) levels

(R2) **participant** @ (n = 391) levels

The two fixed factors were fully crossed, yielding six distinct conditions: *square |0 degrees*, *square |45 degrees*, *square |90 degrees*, *triangular |0 degrees*, *triangular |45 degrees*, *triangular |90 degrees*. Participants were nested within condition, and questions were fully crossed with condition. Thus, each participant was randomly assigned to one of the six (factorial) conditions, in which they completed all questions.

4.6.1.3 Materials & Procedure

The same task, and procedure were used as described in Study 4A.

4.6.1.4 Analysis

Response Accuracy

To test hypotheses related to response ACCURACY we fit mixed logistic regression models in R using the `lme4` package. For contrast coding, the default treatment (dummy) coding scheme was used with the following reference categories:

- response variable ACCURACY : level *0:incorrect* as reference
- predictor factor **shape**: level *square (control)* as reference
- predictor factor **rotation**: level *0 degrees (control)* as reference

To determine a final model we first defined the maximal random effects structure theoretically justified by the study design (random intercepts for questions and subjects). We then fit the most complex model indicated by the study design, including (main) fixed effects for **shape** and **rotation** scaffold as well as their interaction term, and used a likelihood ratio test to determine if this model was superior to a simpler model including fixed main effect only. Statistical significance of each predictor in the superior model was determined via Wald Chi-Square tests, and all reported p-values are for non-directional tests with a decision threshold $\alpha = 0.05$.

4.6.2 Results

4.6.2.1 Total Score

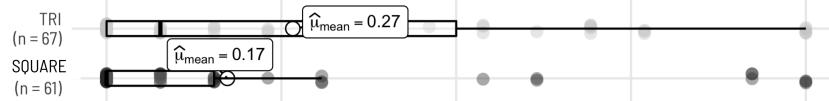
To explore the effect of **shape** and **rotation** on TM graph reading performance, we start by describing the distribution of TOTAL SCORE. Across all conditions, TOTAL SCORE ranged from 0 to 100 with a mean of 33%. In Figure 4.13 we see that participant level accuracy on the graph comprehension task follows the same pattern of behaviour observed in Study 4C: scores are slightly higher for *triangular* (vs) *orthogonal* shaped axes. Most dramatically, we see that across both graph shapes, total score is substantially higher when the graphs were rotated in space by *45 degrees*.

4.6.2.2 Question Accuracy

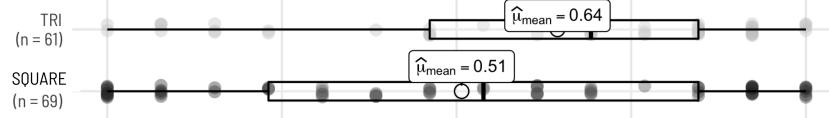
To explore the effects of graph **shape** and **rotation** on ACCURACY, we fit a mixed effects logistic regression model with random intercepts for subjects and questions, with **shape** and **rotation** as fixed effects. A likelihood ratio test indicates that a model including these main effects explains significantly more variance in the data than an intercepts-only baseline model ($\chi^2(3, 6) = 109.40, p < 0.001$). We also fit a model including an interaction term between **shape** and **rotation**, however a likelihood ratio test indicated that adding the interaction term did not improve model fit ($\chi^2(6, 8) = 0.04, p = 0.98$). Therefore we chose the simple main effects model (with random intercepts) as the final model. The explanatory power of the entire model

STUDY 4D | Distribution of Total Score

ROTATION: 90 degrees



ROTATION: 45 degrees



ROTATION: 0 degrees

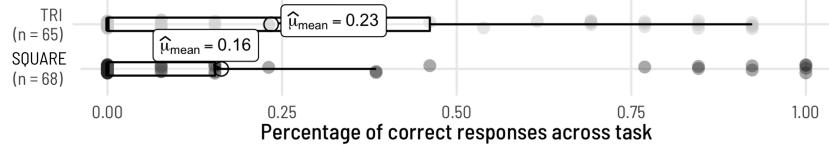


Figure 4.13. Study 4D – Distribution of Total Score. Rotating the both the *triangular* and *orthogonal* shaped graphs by 45 degrees yields a substantial increases in total score.

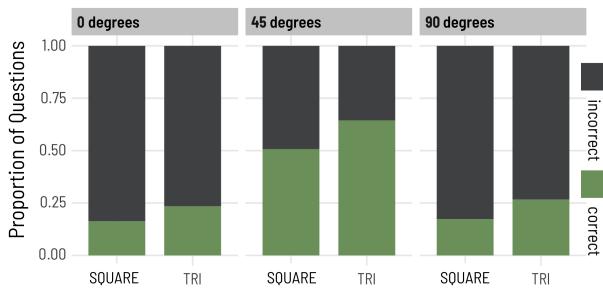
is substantial (*conditional R*² = 0.79) and the part related to the fixed effects (*marginal R*²) explains 21% of variance.

Wald Chi-Square tests revealed significant main effects of both **shape** and ($\chi^2(1) = 10.8, p < 0.001$) **rotation** ($\chi^2(1) = 94.9, p < 0.001$). Model coefficients indicate that across both triangular and orthogonal shaped axes, partially rotating the TM graph (by 45 degrees) increases the odds of a correct response by a factor of 54 ($e^{b_1[45\text{degrees}]} = 54.3, SE = 24.1, p < 0.001$). Across all levels of rotation, shifting the y-axis from orthogonal to triangular increases the odds of a correct response by a factor nearly 3 ($e^{b_1[\text{triangular}]} = 2.8, SE = 0.98, p < 0.01$).

Partially consistent with our (H1) hypothesis, we find that rotation does improve accuracy, but only at the 45 degree rotation level. The model predicts that, for *triangular* shaped graphs, rotating the graph from 0 to 45 degrees increases the probability of a correct response from 5% to 73%. Further rotating the graph to 90 degrees however, results in a probability of only 9%. Post-hoc comparisons indicate that the difference between 0 and 90 degree rotation is not significant (OR = 0.52, SE = 0.23 , p = 0.14).

STUDY 4D | ACCURACY

A Distribution | Question Accuracy



B Model | Probability of Correct Response

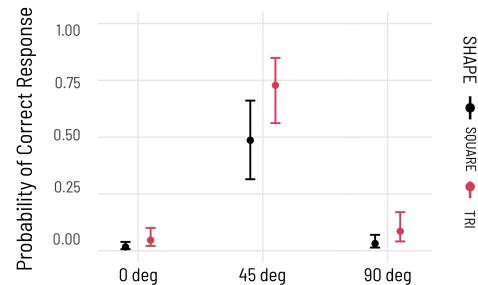


Figure 4.14. Study 4D – Accuracy. A proportional bar chart of raw data shows that the proportion of correct responses is slightly higher for the triangular shape across each rotation level. Rotating the graph by 45 degrees results in a substantial increase in accuracy.

Consistent with our (H2) hypothesis, we find that rotation improves accuracy across both graph axes *shapes*. The model predicts that, for *orthogonal* shaped graphs, rotating the graph from 0 to 45 degrees increases the probability of a correct response from 2% to 49%. Further rotating the graph to 90 degrees however, results in a probability of only 3%. Counter to our (H1) hypothesis post-hoc comparisons indicate that the difference between 0 and 90 degree rotation is not significant (OR = 0.60, SE = 0.26 = 0.25).

Consistent with our (H3) hypothesis, we find that *triangular* shaped y-axis improves accuracy relative to the *orthogonal* shaped axis. Post-hoc comparisons reveal that across all levels of rotation, the triangular shape was significantly better than the orthogonal shape. Model predictions are visualized in Figure 4.14 [B], while parameter estimates and model specification is detailed in Appendix C.4.1.

4.6.3 Study 4D Discussion

In Study 4D we find that across both graph shapes, rotating the graph from 0 to 90 degrees results in a negligible improvement in accuracy, while rotating the graph halfway between results in a substantial improvement in accuracy. These results support our hypothesis that making it more difficult for the reader to mentally project errant lines onto the graph can improve

interpretation accuracy.

What is most promising about this design improvement is that, much like changing the angle of the gridlines (an unsuccessful design change in Study 3C) rotation the graph in space is very salient change, that may be assisting readers in more than one way. First, it draws attention to the fact that the graph is not cartesian. Secondly, it makes the most common mistake more difficult to make, much like the *impasse-structure* in Study 3B. In this case it is not *impossible* to make the mistake, but much more taxing. It is important to note that, much like changing the angle between the x axis and the gridlines (as in Study 4C; scale factor), rotating the graph in space does not change the nature of the coordinate system, or operations for performing interval algebra with it. It could, however, make some of those visual operations more challenging, especially if they require mental projection, such as finding an interval that starts at a time *not* on the hour—requiring the reader to project a diagonal gridline halfway between two existing gridlines. We expect this is an easier projection to make, as the reader can follow the two parallel gridlines. It is not uncommon that we need to project additional gridlines onto a graph, so long as they are oriented consistently with the gridlines as they are drawn.

4.7 General Discussion

Our goal in this chapter was evaluate which if any alterations to the discretionary design features of the TM graph might improve its discoverability (without changing the nature of the graphical formalism). We found that changing the design of the horizontal and diagonal gridlines to either emphasize or obscure the bounding triangle did not improve accuracy (Study 4A). Changing the mark representing an interval from a point to an arrow made a small but statistically significant improvement (Study 4B). Similarly changing the orientation of the y-axis from orthogonal with the x-axis to diagonal (a more spatially efficient design) made a small but reliable improvement in accuracy (Study 4C). And across both this triangular improvement and the square shaped control, partially rotating the graph in space substantially improved accuracy

(Study 4D).

These results support our claim that discretionary design features can be leveraged in ways other than making the graph primary message of a graph more salient; such as making using color or annotations to emphasize a trend or comparison (a second order reading). Or alternatively, by making first order readings more efficient; such as by minimizing visuospatial features unnecessary for supporting inferences) like 3D chart-junk or overly dense grids. Rather, the function of our design changes were to improve discoverability of the coordinate system itself.

4.7.1 Design Implications

The design implications of these results are relevant to both the **invention** of graphical formalisms, as well as the **design** of particular graph instances. With respect to invention, if the formalism being invented is intended to be used by novice readers (i.e. a graph or diagram to communicate with non experts, or the general public) then in addition to computational efficiency, the inventor needs to carefully consider features that might on the surface appear discretionary (such as the shape of a mark, or orientation of the axes) but in fact can serve as cues to discovery. If the new formalism is not intended for general purpose communication (i.e. as was the case for TM graph, designed to support concurrent complex queries over interval relations: a mathematical object), then efficiency can be prioritized over discoverability. *But* for the designer of an instance of the graph, these otherwise discretionary design features are relevant once more as an aid to help (even experts) *learn* to use the new formalism, if they encounter it in a self-directed learning environment (such as a scholarly publication).

4.7.2 Research Implications

What does these result tell us about the graph schema? The relatively small effect size of interventions targeting mark shape and grid design suggest that neither of these features likely lead readers to instantiate a different (non-cartesian) graph schema. Rotating the graph in space,

however, had a relatively large effect. And so it is possible that this rotation might result in a different schema being returned by Pinker's theorized MATCH process. What we cannot be certain of, however, limited by the design of these studies, whether success in TM graph reading (aided or unaided) is the result of a *different* schema being returned (as compared with readers who make the cartesian error), *or* if these successful readers are actively constructing a *new* schema.

We don't know nearly enough about the theorized *general graph schema*. What parameters does it contain, and what operators does it support? We believe one plausible answer to this question, assuming a hierarchical schema structure, is that one that relies on more primitive relations between marks in space, such as those described by Barbara Tversky (2011) in *Visualizing Thought : spatial sequence represents order, dots indicate, lines, connect, boxes contain, arrows lead*. We might think of this as moving up a layer of abstraction from a 'graph' specific schema, to a 'diagram' schema. Imagine, for example, a reader (or computer program) that has no prior knowledge (or operations) for a cartesian coordinate system. If the only operations you have are that: (1) glyphs are meaningfully related in space, and (2) lines define relationships between marks, might you find the TM graph remarkably simple to interpret? *Follow the lines between the glyphs (and their labels).*

Following Pinker's information processing model, it would be ideal if the following processing occurred in the TM graph reading task:

1. The reader encounters the TM graph, and its visuospatial properties are transformed into a visual array.
2. Bottom-up encoding mechanisms transform the visual array into a (propositional) visual description.
3. The MATCH process searches the readers catalogue of graph schemata for a structure appropriate to the visual description. If an appropriate structure is not found, a more *general* (higher order) schema is instantiated.

4. The task question (i.e. conceptual question) is answered via interrogation of the instantiated *general graph schema*.
5. The correct answer (i.e. conceptual message) is derived.

We have substantial evidence of the systematic errors in reading the TM graph, but the design of the present studies does not allow us to differentiate between explanations of how the structure or content of the graph schema influences coordinate system interpretation. The simplest interpretation is that a ‘schema error’ has occurred. If we are to follow Pinker’s theory as an organizing framework, something goes awry in step #3. But we cannot know from this evidence precisely what. Are errors in interpretation a result of instantiating the wrong schema because there *is* a more general schema does not exist, or because we MATCH on a cartesian schema and fail to instantiate the more general form? For orthogonal-interpreters, is a cartesian coordinate schema the most general schema? And if that is the case, how might these readers fare with pie charts, and graphs with polar coordinates? Is the work of correctly reading the graph as simple as the MATCH process instantiating a more general schema, *or* is it the work of constructing a new schema altogether? Are we mistaken in presuming theory in *graph* comprehension can be extended along the murky continuum from forms easily defined as graphs (charts, plots) toward those with more flexible spatial relations: diagrams? To effectively address these questions, we need to move beyond the exploration of the TM coordinate system, and compare performance across multiple coordinate systems, in relation to the more primitive graphic forms we argue might constitute a general graph(ic) schema.

Acknowledgements

This chapter, in part, is currently being prepared for submission for publication of the material. The dissertation author was the primary investigator and author of this material.

Chapter 5

Conclusion

The catalyst for this dissertation was a personal experience: my own encounter with a simple but unconventional graph; my confusion, frustration, and subsequent fascination with the experience of discovering the rules of a new representational system. In one moment I couldn't read this strange graph, and in another, I couldn't imagine not being able to read it. My reflections on this encounter (in the moments and the years thereafter) have left me convinced that although novel representations are by their virtue, rare, they nonetheless offer insight into an important phenomenon that has been heretofore understudied. The systematic errors that humans make when interacting with these forms reveal gaps in our understanding of graph comprehension and the graphical discovery that must precede it, implications for the invention and design of visualizations, as well opportunities for integration of theory from disparate areas of cognitive research.

5.1 Summary of Findings

1. In **Study 1** we observed how readers approach interaction with the novel TM graph. We found that most readers disregard the diagonal gridlines and misinterpret the coordinate system as cartesian. In doing so, readers violate multiple graph reading norms, including:
 - (1) accepting that some questions had no answers, (2) needing to add information (extra lines) to the graph to solve the problems, (3) needing to read past the end of the numerical

axes, and (4) accepting the presence of information on the graph with no meaning. During a debrief interview, after discovering (or being instructed on the function of) the coordinate system, participants believed that minor additions to the graph (text or image-based instructions) would improve discoverability for future participants.

2. In **Study 2** we evaluated four explicit scaffolds inspired by the designs produced in Study 1. We found that text and image instructions improved graph interpretation for some but not all subjects, and that an interactive image condition was by far the most effective. When asked to produce a TM graph for a small dataset, most readers produced accurate graphs, even if they had not done so during the graph reading task.
3. In **Studies 3A-C** we explored the extent to which graph discovery can be construed as an insight problem. We found that imposing an impasse structure for participants reading the TM graph significantly improved interpretation, though the intervention was less effective than either explicit static or interactive image scaffolds. In a follow up study measuring Working Memory Capacity, we found that WMC explained some of the variance introduced by individual differences, specifically that only individuals with high working memory capacity were able to take advantage of an *impasse structure* to correctly interpret the graph.
4. In **Studies 4A-D** we explored the role of discretionary design features on coordinate system discoverability, finding that changing the data-marks and axis orientation yielded a minor but statistically significant improvements in interpretation, but changing the design of the gridlines did not. Most impressively, partially rotating the graph in space substantially improved discoverability of the coordinate system.

Taken together, the results of these studies indicate that **discovering the formalism of an unconventional graph is much harder than we expect** and that performance is characterized by a combination of **systematic errors** and **individual differences**. Some readers are able to

successfully discover the rules of a new representational system with little or no guidance, and others fail to do so, even in the face of substantial scaffolding. Most readers are not aware that they are misreading the graph. Prior knowledge of the cartesian coordinate system is **incredibly difficult to overcome**.

5.2 Implications for Visualization Design

A naive interpretation of our results might be that the TM graph is simply *a very bad graph*. How can the graph be any good, if it is so very hard to discover how to read? This visualization researcher, however, insists that graphs are neither universally good nor bad, but rather, more or less *effective*, for a particular set of data, for a particular task, for a particular audience, in a particular communicative context. For trained readers performing interval calculus, the computational efficiency of the TM graph relative to a more conventional representation of time intervals has been empirically demonstrated. Once you know how to read it, its simplicity (and elegance) is self-evident.

The design implications of these results are relevant to both the **invention** of graphical formalisms, and the **design of scaffolding** and instructions to support communication via graph instances. With respect to invention, if the formalism being invented is intended to be used by novice readers (i.e. a graph or diagram to communicate with non experts, or in the context of learning) then in addition to computational efficiency, the inventor should consider otherwise discretionary features (such as the shape of a mark, or orientation of the axes) as cues to discovery. If the new formalism is not intended for general non-expert communication then computational efficiency can be prioritized over discoverability in making design decisions.

But the designer of instructional scaffolds (for either general purpose communication or the introduction of a novel form to an expert audience) should again carefully consider which design features are in fact discretionary. Our results suggest that *more is more* when it comes to scaffolding, and the strategic combination of text, image and impasse-structure scaffolds can

be incredibly effective at facilitating self directed learning. In the absence of explanations for individual differences in the ability to discover a novel formalism, my design recommendation is to leverage multiple encodings of instructions, to reach as many readers in as many ways as possible.

5.3 Implications for Research

The Source of Individual Differences. Through the empirical studies in this dissertation we've gathered substantial evidence of the magnitude and variety of errors readers make with the TM graph. We've also seen tremendous individual differences in performance. Although several of the interventions aimed at improving discoverability were effective (i.e. text instructions, static and interactive images, impasse structure, mark design, shape design, rotation design), in all cases there were readers who persisted in a cartesian misinterpretation: scores never reached ceiling. Similarly, some readers were able to discover the interval coordinate system with ease. In Study 3C we explored working memory capacity as one source of these individual differences, but found only that WMC was predictive of interpretation in the presence of the *impasse structure*. Working memory capacity did not explain task success or failure in the control condition, suggesting that other individual differences are relevant to the kind of representational restructuring thought to be required to reconceptualize a coordinate system. Coordinate systems are fundamentally *spatial*, but also *mathematical objects*. And thus we suspect the most likely candidates for explanatory variables are capabilities in: (1) visuospatial reasoning (a construct as nuanced as working memory), (2) formal math education, (specifically trigonometry, and courses that emphasize the function concept, function graphing, or non-euclidean geometry), (3) and a broader range of skills DiSessa refers to as *meta-representational competency* (A. diSessa, 2004; A. diSessa, Hammer, and Sherin, 1991; A. A. diSessa and Sherin, 2000).

Explanatory Power of the Graph Schema. As presently expressed in the research literature, the *graph schema* is a catch-all term to describe prior knowledge of graph types, and how

they work. Many failures to read a graph as it is designed are described as ‘schema errors’ (L. Padilla, Castro, and Hosseinpour, 2021; L. M. Padilla, Creem-Regehr, Hegarty, and Stefanucci, 2018). As a framework for organizing our knowledge about types of computations involved in comprehending a graph, and the sequence in which some of these events occur, Pinker’s theory of graph comprehension is extraordinarily useful (and continues to be widely cited). But the construct of the graph schema in particular is underspecified, and difficult to empirically evaluate (see Ratwani and Trafton, 2008). To offer useful predictions as to the nature of errors that occur in specific situations, the construct of the graph schema needs to be further elaborated, or replaced with a more powerful explanatory mechanism. I suggest that one potential way forward is to explore a phenomenon that the graph schema fails to sufficiently explain (such as graph discovery) through the multiple theoretical lenses, and explore what explanatory heavy lifting constructs in other theoretical frames might offer. This approach has already proven fruitful in the elaboration of theory describing decision making with visualizations (see L. M. Padilla, Creem-Regehr, Hegarty, and Stefanucci, 2018), evolving out of an integration of Pinker’s Theory of Graph Comprehension with dual-process theories of judgement and decision making. Specifically with respect to graph *discovery*, in Chapter 3 we explored the applicability of constructs from the problem solving literature, including *negative transfer*, *insight* and *impasse*. Exploring how the problem solving literature accounts for the kind of problem restructuring required to reach a correct solution, and comparing this to how the graph comprehension literature addresses the same requirement in the graphical context, is a step toward theoretical integration.

5.4 A Personal Reflection, and Future Work

When I started my PhD I was warned by a sage upperclassman that the work would never be finished, that I would graduate knowing mostly more about how much I didn’t know, and that I would have some degree of disdain for my dissertation topic. Two of these predictions came to pass. I would have liked for this research to yield more answers than questions; that

was certainly the goal. But as I reflect on the results, the methods, the phenomenon, and the research process, I am both heartened and excited by the reality that there is so much left to be discovered. I have a newfound appreciation for the value of descriptive research. When I watch a mouse-cursor replay of a TM graph reader *fail* to discover its coordinate system I continue to vacillate—like an ambiguous figure —between incredulity at their inability to *simply* make use of the salient diagonal gridlines, and certainty that the cartesian interpretation is of course, the most rational behaviour. Perhaps also a wisp of empathy, borne of my own struggle, and wish that I had been one of those TM graph discoverers. But even as an information designer and graph comprehension researcher, I wasn’t. I looked up the instructions. And that makes me want to know *why* given so much knowledge (and motivation!) even I found the task so difficult. And will I find the task of discovering a *different* novel coordinate system similarly difficult? Will either of these difficulties bear on my ability to design or interpret *diagrams*? And will the answer to these questions tell me anything about my ability to invent new representational systems, or design scaffolds for others to discover their efficiencies? Perhaps most importantly, why do I find only breadcrumbs toward satisfactory answers, spread across seemingly disparate areas of cognitive research? And is that truly a problem, or rather a marvellous opportunity?

Appendix A

Study 1 & 2 Supplementary Material

A.1 Study 1 & 2 — Materials

The following six pages include the scenario instructions, questions, and stimulus graphs used in Study 1 and Study 2, as referenced in Chapter 2.

Axis Hackathon

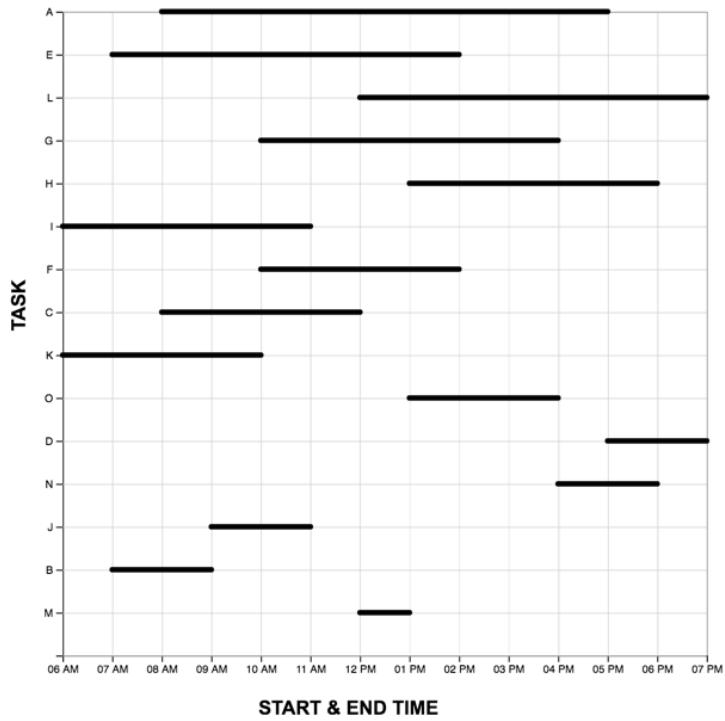
Jane is the Project Manager for team Axis in a health care hackathon. A hackathon is an event, typically lasting 24 hours, where designers and developers come together to build a product. Jane is responsible for keeping her team on track to finish their ambitious project in only 24 hours. She has compiled a list of tasks and organized them into a schedule. Now she needs to assign the tasks to her team.



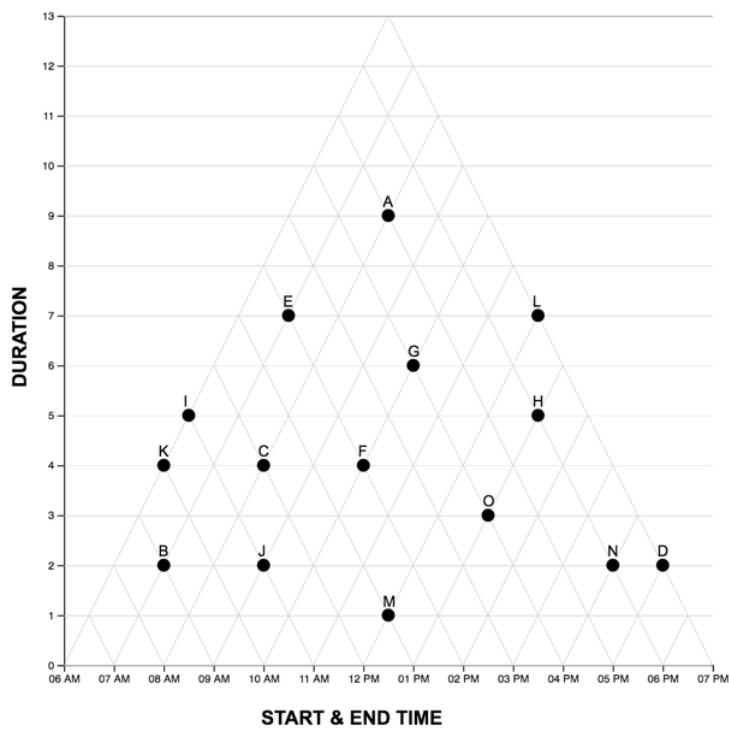
Help Jane organize the schedule and make assignments for her team.

QUESTION	<relation>
Which tasks are scheduled to start at 1pm?	start
When does task G end?	end
Jane wants to divide tasks four hours or longer into smaller tasks. Which should she consider subdividing?	duration
Jane would like her teammates to check in with her at the midway point (half way through) each of their tasks. What is the first time during the day that two teammates will update Jane at the same time?	midpoint
Of all the tasks that overlap with H, what is the earliest start time?	start+overlap
All of the tasks that last two hours involve testing prototypes. When does the second round of prototype testing begin?	start+duration
Which tasks are performed before N and do not overlap with K?	before+overlap
Which tasks do not overlap with I or D?	overlap+overlap
Which room is reserved for the whole time that room H is reserved, and also overlaps with room D?	during+overlap
Which tasks under four hours are scheduled during the same time span as task L?	contain+duration
How many times during the day do two tasks start at the same time?	starts-with
What tasks are scheduled to end at 4PM?	finishes-with
Which task(s) takes place entirely during task G?	during
Jane wants to assign Alex to tasks K and N. What other task(s) can Jane assign to Alex so that he is working from 6AM to 6PM?	meets
Which task starts with F and ends with O?	starts+finishes
Which task starts sometime after F and is one hour shorter?	before+duration

Axis Scenario LM Graph



Axis Scenario TM Graph



Longmire Inn & Conference Center

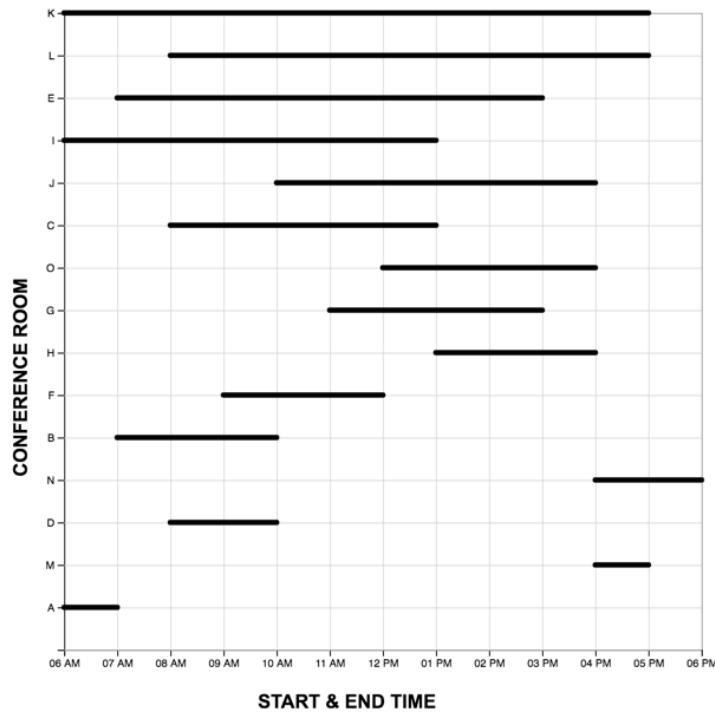


Teresa is the Manager of Event Operations for the Longmire Inn & Conference Center. The property has meeting rooms and event spaces that can be reserved for conferences and parties. Tomorrow is a busy day for Teresa because every conference room is reserved at some point during the day.

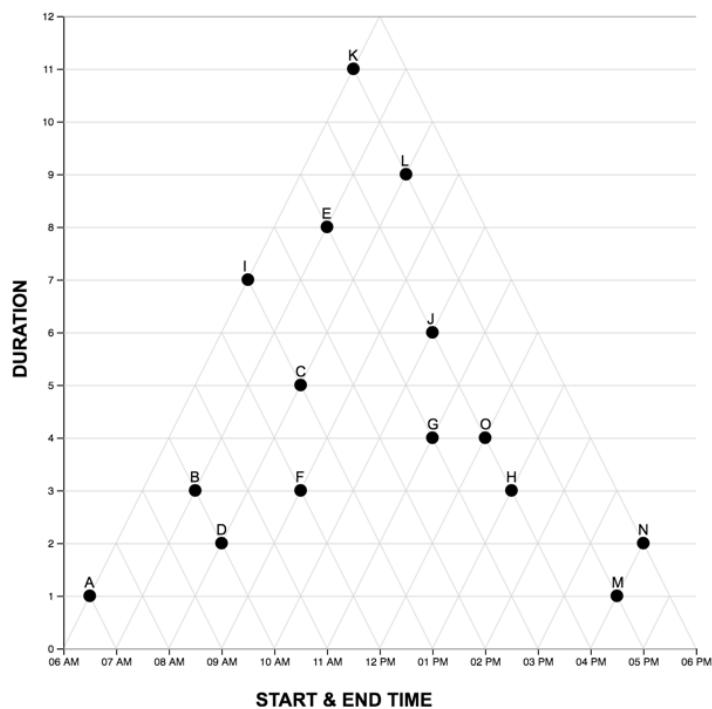
Help Teresa review the event schedule so she can organize her staff.

QUESTION	<relation>
Which room reservations start at 8:00 AM?	start
Which room reservations end after 3:00 PM?	end
Which rooms are reserved for 4 hours?	duration
Guests in the room J have requested a snack break midway through their event. At what time do the staff need to have the snacks ready?	midpoint
What is the earliest reservation that does not overlap with F?	start+overlap
Events in the morning (end at noon or earlier) that last less than 3 hours are eligible for a discounted rate. Which rooms are eligible?	start+duration
Which reservations are before N and do not overlap with H?	before+overlap
Which reservations do not overlap with D or H?	overlap+overlap
Which room is reserved for the whole time that room H is reserved, and also overlaps with room D?	during+overlap
Which 2 rooms are reserved for the same number of hours, and are entirely contained by the time that room J is reserved?	contain+duration
When two or more reservations start at the same time, Teresa needs to get help from the housekeeping staff to prepare the rooms. For what reservation times does Teresa need extra staff?	starts-with
When three or more reservations END at the same time, Teresa needs to get help from the housekeeping staff to prepare the rooms. For what reservation times does Teresa need extra staff?	finishes-with
Teresa's event manager Liam is personally attending to the reservation in conference room E for a high priority client. Which reservations can Liam not set up while he is attending to room E?	during
The company reserving conference room F have also reserved a second room. Immediately after they finish with room F, to which room will they relocate?	meets
Which task starts with D and ends with I?	starts+finishes
Which reservation starts before room J and is one hour shorter?	before+duration

Longmire Scenario LM Graph



Longmire Scenario TM Graph



Jones Clinical Laboratory

James is the Lab Manager for Dr. Jones' clinical research lab. James is responsible for managing the busy schedule of patients that come into the lab to participate in clinical trials. Four research studies are currently underway, and James is overwhelmed by the task of scheduling appointments! It is very important that each patient be scheduled for the right clinical study, and for the correct number of hours.



You are going to help James check his schedule for errors by drawing a graph of his patient scheduling data.

Drawing Instructions

You are going to help James check his schedule for errors by drawing a graph of his patient scheduling data.

The experimenter has provided you with a pencil and sheet of graph paper. Using these materials, please draw a **triangular interval graph**. The triangular interval graph is one of the graphs you used in the previous activities (the one with the diagonal grid lines). Please plot each of the appointments listed on the data table on the graph.

When you are finished, please check that you have:

- Write [subject code] on the "Code" line
- Write your session on the "Session" line
- Put a title on your graph
- Put a title on each of the axes
- Labelled the tick marks on the axes
- Labelled each data point (as needed)

Patient #	Trial #	Start Time	End Time
A1	1	7:00 AM	9:00 AM
D2	2	8:00 AM	9:30 AM
B1	1	9:00 AM	11:00 AM
G3	3	9:00 AM	12:00 PM
I4	4	11:00 AM	12:00 PM
C1	1	12:00 PM	3:00 PM
E2	2	1:00 PM	2:30 PM
F3	3	1:00 PM	4:00 PM
H3	3	3:00 PM	6:00 PM
J4	4	4:00 PM	5:00 PM

Code: _____ Session: _____

A.2 Study 2 — Supplemental Results

A.2.1 Results : The Need for Scaffolding

Study 2 | Linear Graph Accuracy is consistent across conditions

$\chi^2_{\text{Kruskal-Wallis}}(4) = 0.35, p = 0.99, \hat{\varepsilon}_{\text{ordinal}}^2 = 1.12e-03, \text{CI}_{95\%} [2.57e-03, 1.00], n_{\text{obs}} = 316$

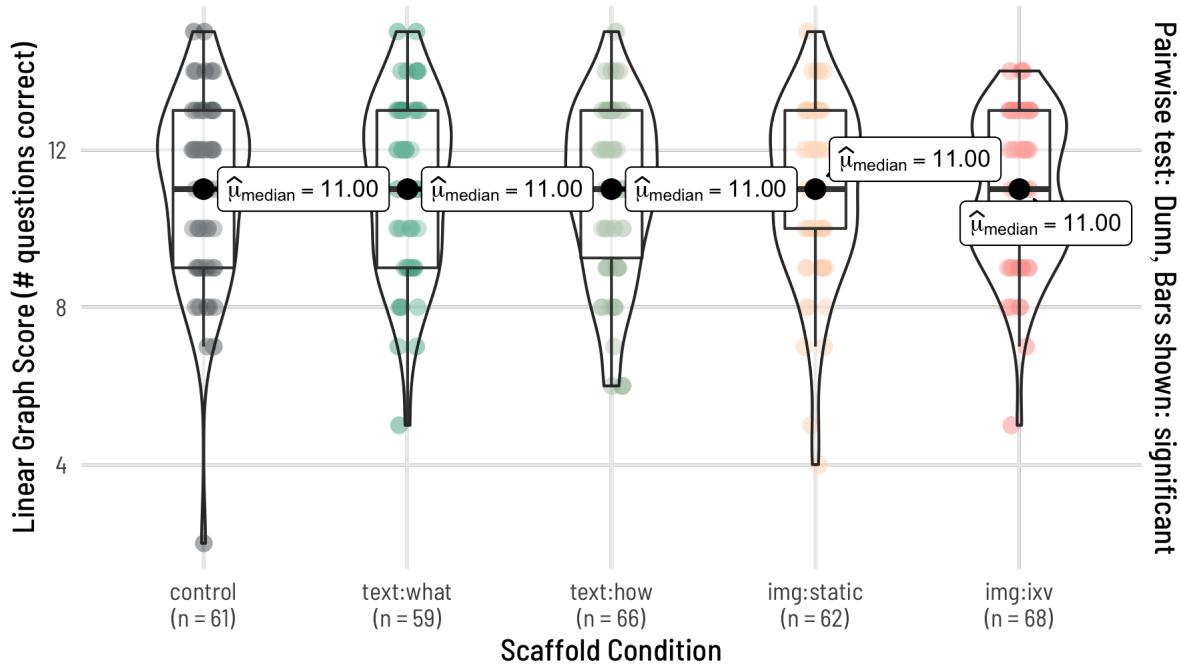


Figure A.1. Study 2 (Results) Linear Model Results by Scaffold. Median response score by condition, for the Linear Model graph. Scores do not differ across scaffold conditions, indicating that scaffolding is not needed for the Linear Model graph. *See section 2.4.2.2*

Study 2 | Without scaffolding, higher accuracy with Linear Model

$V_{\text{Wilcoxon}} = 1490.00, p = 1.33e-07, \hat{r}_{\text{biserial}}^{\text{rank}} = 0.80, \text{CI}_{95\%} [0.67, 0.88], n_{\text{pairs}} = 61$

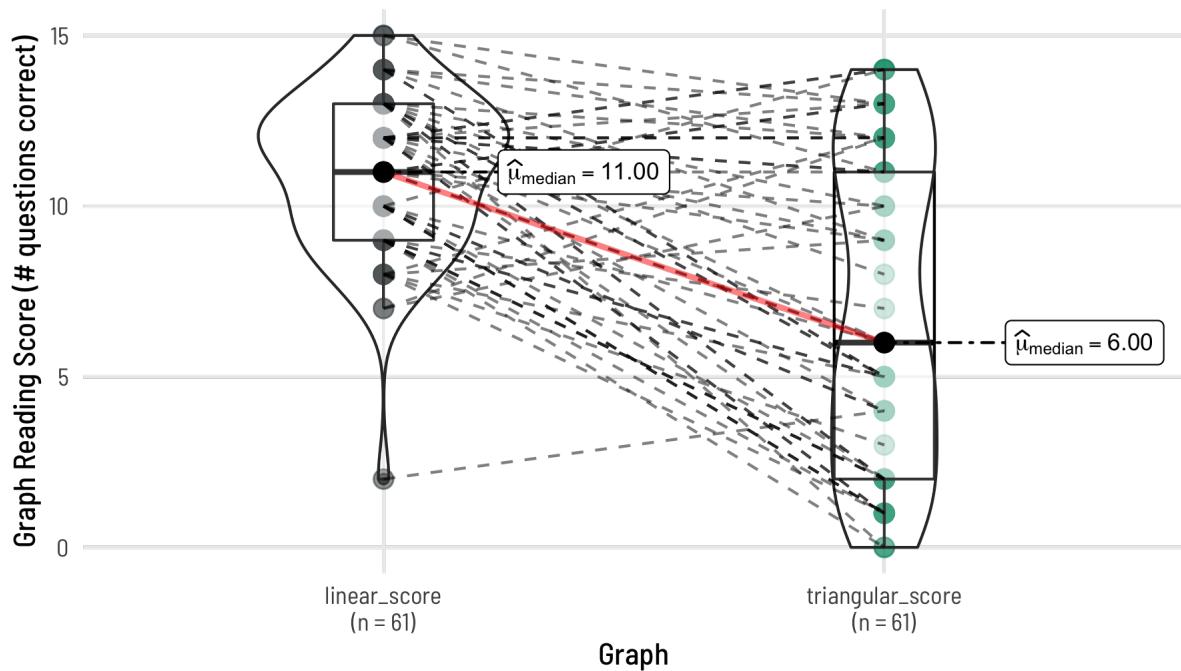


Figure A.2. Study 2 (Results) Paired Scores by Graph Block. Median response score for each graph (in the control condition) indicates that most readers are significantly more accurate with the Linear Model graph than the Triangle Model graph, supporting the hypothesis that the Triangle Model requires scaffolding to support discoverability. See section 2.4.2.2

Appendix B

Study 3A — 3C | Supplementary Material

B.1 Study 3A — Lab | Supplementary Material

B.1.1 Accuracy Model

A Mixed Effects Logistic Regression Model indicates that **implicit scaffold** condition has a significant effect on response ACCURACY (*see Section 3.3.2.2*).

$$\widehat{\text{accuracy}}_i \sim \text{Binomial}(n = 1, \text{prob}_{\text{accuracy}=\text{correct}} = \hat{P})$$

$$\log \left[\frac{\hat{P}}{1 - \hat{P}} \right] = -5.24 \alpha_{j[i], k[i]}$$

$$\alpha_j \sim N(4.13 \gamma_1^\alpha (\text{condition}_{\text{impasse}}), 4.58), \text{ for subject } j = 1, \dots, J$$

$$\alpha_k \sim N(0, 0.57), \text{ for q k = 1, \dots, K}$$

Table B.1. Study 3A (Lab) | Question Accuracy

Mixed Logistic Regression via lme4 (GLMER)

	odds ratios			(log odds)		
	Est.	2.5 %	97.5 %	Est.	2.5 %	97.5 %
(Intercept)	0.01 ***	0.00	0.04	-5.24 ***	-7.20	-3.28
Condition[impasse]	61.90 ***	7.21	531.75	4.13 ***	1.97	6.28
SD (Intercept subject)	97.79			4.58		
SD (Intercept q)	1.77			0.57		
SD (Observations)	2.72			1.00		

Model ACCURACY ~ Condition + (1|subject) + (1|q)

$n = 1638$ $R^2(\text{Conditional}) = 0.89$. $R^2(\text{Marginal}) = 0.15$

* p < 0.05, ** p < 0.01, *** p < 0.001

B.1.2 Interpretation Model

A Bayesian Mixed Effects Multinomial Regression Model indicates that evidence for a reliable effect of **implicit scaffold** condition on the probability of each transitional response INTERPRETATION. Readers in the *impasse* condition produce more *other*, *angular* and *triangular* responses than participants in the *non-impasse* control condition (see Section 3.3.2.4).

Table B.2. Study 3A (Lab) | Question Interpretation

Bayesian Mixed Multinomial Regression via brms (family = categorical)

Interpretation	Parameter	(Odds Ratio)	CI_low	CI_high	pd	%_in_ROPE
other	(Intercept)	0.04	0.02	0.10	1	0
other	Condition[impasse]	12.13	6.29	25.24	1	0
angular	(Intercept)	0.01	0.00	0.03	1	0
angular	Condition[impasse]	11.48	3.95	37.67	1	0
triangular	(Intercept)	0.02	0.00	0.08	1	0
triangular	Condition[impasse]	33.90	6.22	211.18	1	0

Model INTERPRETATION ~ *condition* + (1|*subject*) + (1|*q*)

Bayes Factor = 1.4e+14 (against random intercepts only model)

B.2 Study 3A — Online Replication | Supplementary Material

B.2.1 Accuracy Model

A Mixed Effects Logistic Regression Model indicates that **implicit scaffold** condition has a significant effect on response ACCURACY (*see Section 3.4*).

$$\begin{aligned} \widehat{\text{accuracy}}_i &\sim \text{Binomial}(n = 1, \text{prob}_{\text{accuracy}=\text{correct}} = \hat{P}) \\ \log \left[\frac{\hat{P}}{1 - \hat{P}} \right] &= -3.54 \alpha_{j[i], k[i]} \\ \alpha_j &\sim N(2.2 \gamma_1^\alpha (\text{condition}_{\text{impasse}}), 3.86), \text{ for subject } j = 1, \dots, J \\ \alpha_k &\sim N(0, 0.74), \text{ for q k = 1, \dots, K} \end{aligned}$$

Table B.3. Study 3A (Replication) | Question Accuracy

Mixed Logistic Regression via lme4 (GLMER)

	odds ratios			(log odds)		
	Est.	2.5 %	97.5 %	Est.	2.5 %	97.5 %
(Intercept)	0.03 ***	0.00	0.18	-3.54 ***	-5.39	-1.69
Condition[impasse]	9.06 *	1.04	79.08	2.20 *	0.04	4.37
SD (Intercept subject)	47.50			3.86		
SD (Intercept q)	2.09			0.74		
SD (Observations)	2.72			1.00		

Model ACCURACY ~ Condition + (1|subject) + (1|q)

$n = 923$ $R^2(\text{Conditional}) = 0.83$ $R^2(\text{Marginal}) = 0.06$

* p < 0.05, ** p < 0.01, *** p < 0.001

B.2.2 Interpretation Model

A Bayesian Mixed Effects Multinomial Regression Model indicates strong evidence for a reliable effect of **implicit scaffold** condition on the probability of each transitional response INTERPRETATION. Readers in the *impasse* condition produce more *other*, *angular* and *triangular* responses than participants in the *non-impasse* control condition (see Section 3.4).

Table B.4. Study 3A (Replication) | Question Interpretation

Bayesian Mixed Multinomial Regression via brms (family = categorical)

Interpretation	Parameter	(Odds Ratio)	CI_low	CI_high	pd	%in_ROPE
other	(Intercept)	0.14	0.05	0.37	1.00	0.00
other	Condition[impasse]	4.57	2.27	9.52	1.00	0.00
angular	(Intercept)	0.04	0.01	0.12	1.00	0.00
angular	Condition[impasse]	3.69	1.29	10.91	0.99	0.00
triangular	(Intercept)	0.07	0.01	0.36	1.00	0.00
triangular	Condition[impasse]	8.82	1.04	74.73	0.98	0.01

Model INTERPRETATION \sim condition + (1|subject) + (1|q)

Bayes Factor = 40877 (against random intercepts only model)

B.3 Study 3B — Supplementary Material

B.3.1 Accuracy Model

A Mixed Effects Logistic Regression Model indicates that **implicit** and **explicit scaffold** conditions have an positive, additive effect on response ACCURACY with an interaction at the ceiling scores in the *interactive-image* conditions. (see Section 3.5.2.2).

$$\widehat{\text{accuracy}}_i \sim \text{Binomial}(n = 1, \text{prob}_{\text{accuracy}=\text{correct}} = \hat{P})$$

$$\log \left[\frac{\hat{P}}{1 - \hat{P}} \right] = -3.7 \alpha_{j[i], k[i]}$$

$$\begin{aligned} \alpha_j &\sim N \left(2.87 \gamma_1^\alpha (\text{IMPLICIT}_{\text{impasse}}) \right. \\ &+ 4.63 \gamma_2^\alpha (\text{EXPLICIT}_{\text{img}}) + 6.81 \gamma_3^\alpha (\text{EXPLICIT}_{\text{ixn}}) \\ &- 1.44 \gamma_4^\alpha (\text{EXPLICIT}_{\text{img}} \times \text{IMPLICIT}_{\text{impasse}}) \\ &\left. - 2.59 \gamma_5^\alpha (\text{EXPLICIT}_{\text{ixn}} \times \text{IMPLICIT}_{\text{impasse}}) \right), 3.06, \text{ for subject } j = 1, \dots, J \\ \alpha_k &\sim N(0, 0.82), \text{ for q k = 1, \dots, K} \end{aligned}$$

Table B.5. Study 3B (Explicit vs Implicit) | Question Accuracy

Mixed Logistic Regression via lme4 (GLMER)

	odds ratios			(log odds)		
	Est.	2.5 %	97.5 %	Est.	2.5 %	97.5 %
(Intercept)	0.02 ***	0.0	0.1	-3.70 ***	-4.8	-2.6
Implicit[impasse]	17.55 ***	4.9	63.1	2.87 ***	1.6	4.1
Explicit[img]	102.73 ***	27.7	380.7	4.63 ***	3.3	5.9
Explicit[ixn]	909.54 ***	227.6	3634.8	6.81 ***	5.4	8.2
Implicit[impasse]:Exp[img]	0.2	0.04	1.4	-1.44	-3.2	0.3
Implicit[impasse]:Exp[ixn]	0.1 **	0.01	0.4	-2.59 **	-4.4	-0.8
SD (Intercept subject)	21.42			3.06		
SD (Intercept q)	2.27			0.82		
SD (Observations)	2.72			1.00		

Model ACCURACY ~ IMPLICIT * EXPLICIT + (1|subject) + (1|q)

n = 4849 R²(Conditional) = 0.83 R²(Marginal) = 0.32

* p < 0.05, ** p < 0.01, *** p < 0.001

B.3.2 Interpretation Model

A Bayesian Mixed Effects Multinomial Regression Model indicates strong evidence for a reliable effect of CONDITION on the probability of each transitional Response Interpretation. Readers in the Impasse condition produce more *unknown/uncertain*, *angular* and *triangular* responses than participants in the control condition (*see Section 3.4*).

Table B.6. Study 3B (Explicit vs Implicit) | Question Interpretation

Bayesian Mixed Multinomial Regression via brms (family = categorical)

Interpretation	Factor	(Odds Ratio)	CIlow	CIhigh	pd	%in_ROPE
other	(Intercept)	0.06	0.03	0.13	1	0
other	Implicit[impasse]	6.92	4.64	10.58	1	0
other	Explicit[image]	4.09	2.66	6.61	1	0
other	Explicit[interactive]	3.25	1.95	5.39	1	0
angular	(Intercept)	0.01	0.00	0.03	1	0
angular	Implicit[impasse]	8.18	4.59	14.66	1	0
angular	Explicit[image]	6.39	3.28	12.33	1	0
angular	Explicit[interactive]	6.78	3.44	13.79	1	0
triangular	(Intercept)	0.06	0.02	0.16	1	0
triangular	Implicit[impasse]	9.96	4.85	21.66	1	0
triangular	Explicit[image]	68.29	26.70	173.35	1	0
triangular	Explicit[interactive]	307.85	116.73	889.41	1	0

Model INTERPRETATION \sim IMPLICIT + EXPLICIT + (1|subject) + (1|q)

Bayes Factor = 5.03e+129 (against random intercepts only model)

B.4 Study 3C — Supplementary Material

B.4.1 Accuracy Model

A Mixed Effects Logistic Regression Model indicates that CONDITION has a significant positive main effect on Question ACCURACY and a significant interaction effect with OSPAN (working memory capacity), such that high working memory participants perform significantly better in the impasse condition (*see section 3.6.2.2*)

$$\widehat{\text{accuracy}}_i \sim \text{Binomial}(n = 1, \text{prob}_{\text{accuracy}=\text{correct}} = \hat{P})$$

$$\log \left[\frac{\hat{P}}{1 - \hat{P}} \right] = -6.83 \alpha_{j[i], k[i]}$$

$$\begin{aligned} \alpha_j &\sim N(2.35 \gamma_1^\alpha (\text{condition}_{\text{impasse}}) \\ &- 0.67 \gamma_2^\alpha (\text{ospan}_{\text{high-memory}}) \\ &+ 4.84 \gamma_3^\alpha (\text{ospan}_{\text{high-memory}} \times \text{condition}_{\text{impasse}}), 5.59), \text{ for subject } j = 1, \dots, J \end{aligned}$$

$$\alpha_k \sim N(0, 1.09), \text{ for q k = 1, \dots, K}$$

Table B.7. Study 3C (Working Memory) | Question Accuracy

Mixed Logistic Regression via lme4 (GLMER)

	odds ratios			(log odds)		
	Est.	2.5 %	97.5 %	Est.	2.5 %	97.5 %
(Intercept)	0.00 ***	0.00	0.02	-6.83 ***	-9.70	-3.96
Condition[impasse]	10.46	0.46	240.44	2.35	-0.79	5.48
ospan[high-memory]	0.51	0.03	8.47	-0.67	-3.47	2.14
condition[impasse]*ospan[high-memory]	127.05 *	1.53	10549.66	4.84 *	0.43	9.26
SD (Intercept subject)	269.01			5.59		
SD (Intercept q)	2.97			1.09		
SD (Observations)	2.72			1.00		

Model Accuracy ~ Condition * OSPAN + (1|subject) + (1|q)

$n = 1729$ $R^2(\text{Conditional}) = 0.92$ $R^2(\text{Marginal}) = 0.18$

* p < 0.05, ** p < 0.01, *** p < 0.001

B.4.2 Interpretation Model

A Bayesian Mixed Effects Multinomial Regression Model indicates mixed evidence for the interaction of CONDITION with OSPAN (working memory capacity). The model predicts similar probabilities for other, and angular interpretations by high vs. low working memory participants, with only a main effect of impasse CONDITION increasing the probability of these transitional interpretations. It is only the (correct) triangular interpretation for which we have evidence for a reliable interaction between OSPAN and CONDITION (see Section 3.6.2.3).

Table B.8. Study 3C (Working Memory) | Question Interpretation

Bayesian Mixed Multinomial Regression via brms (family = categorical)						
Interpretation	Factor	(Odds Ratio)	CIlow	CIhigh	pd	%in_ROPE
other	(Intercept)	0.14	0.06	0.31	1.00	0.00
other	Condition[impasse]	6.13	3.42	11.53	1.00	0.00
other	OSPA[Nhigh]	0.83	0.44	1.56	0.72	0.39
other	Condition:OSPA[N]	0.93	0.38	2.25	0.56	0.32
angular	(Intercept)	0.02	0.00	0.07	1.00	0.00
angular	Condition[impasse]	6.39	2.21	20.08	1.00	0.00
angular	OSPA[Nhigh]	1.18	0.37	3.74	0.61	0.25
angular	Condition:OSPA[N]	0.72	0.16	3.41	0.66	0.17
triangular	(Intercept)	0.01	0.00	0.05	1.00	0.00
triangular	Condition[impasse]	11.64	1.27	107.34	0.99	0.00
triangular	OSPA[Nhigh]	1.07	0.11	9.76	0.52	0.13
triangular	Condition:OSPA[N]	15.73	0.89	249.91	0.97	0.01

Model Interpretation \sim condition * ospan + (1|subject) + (1|q)

Bayes Factor = 2.7 (against random effects only model)

Appendix C

Study 4A — 4D | Supplementary Material

C.1 Study 4A — | Supplementary Material

C.1.1 Accuracy Model

A Mixed Effects Logistic Regression Model indicates that neither altering the gridlines of the TM graph to emphasize the valid interval space (*sparse design*) nor to emphasize the diagonal grid (*grid design*) significantly improves ACCURACY. (see Section 4.3.2.2).

$$\widehat{\text{accuracy}}_i \sim \text{Binomial}(n = 1, \text{prob}_{\text{accuracy}=\text{correct}} = \hat{P})$$

$$\log \left[\frac{\hat{P}}{1 - \hat{P}} \right] = -9.68 \alpha_{j[i], k[i]}$$

$$\alpha_j \sim N \left(0.05 \gamma_1^\alpha (\text{GRIDLINES}_{\text{sparse}}) - 0.52 \gamma_2^\alpha (\text{GRIDLINES}_{\text{grid}}), 8.68 \right),$$

for subject $j = 1, \dots, J$

$$\alpha_k \sim N(0, 0.84), \text{ for } q \ k = 1, \dots, K$$

Table C.1. Study 4A | Question Accuracy

Mixed Logistic Regression via lme4 (GLMER)

	odds ratios			(log odds)		
	Est.	2.5 %	97.5 %	Est.	2.5 %	97.5 %
(Intercept)	0.00 ***	0.00	0.00	-9.68 ***	-11.16	-8.19
GRIDLINES[sparse]	1.06	0.26	4.21	0.05	-1.33	1.44
GRIDLINES[grid]	0.59	0.15	2.31	-0.52	-1.88	0.84
SD (Intercept subject)	5907.01			8.68		
SD (Intercept q)	2.31			0.84		
SD (Observations)	2.72			1.00		

Model ACCURACY \sim GRIDLINES + (1|subject) + (1|q)

n = 5889 R²(Conditional) = 0.96 R²(Marginal) = 0

* p < 0.05, ** p < 0.01, *** p < 0.001

C.2 Study 4B — | Supplementary Material

C.2.1 Accuracy Model

A Mixed Effects Logistic Regression Model indicates that changing the mark that represents the temporal intervals from a *point* to an *arrow* (thus emphasizing the relationship between the point and the diagonal gridlines) significantly improves TM graph reading ACCURACY (*see Section 4.4.2.2*).

$$\widehat{\text{accuracy}}_i \sim \text{Binomial}(n = 1, \text{prob}_{\text{accuracy}=\text{correct}} = \hat{P})$$

$$\log \left[\frac{\hat{P}}{1 - \hat{P}} \right] = -6.79 \alpha_{j[i], k[i]}$$

$$\alpha_j \sim N \left(1.55 \gamma_1^\alpha (\text{MARK}_{\text{cross}}) + 3.46 \gamma_2^\alpha (\text{MARK}_{\text{arrow}}), 5.07 \right),$$

for subject j = 1, ..., J

$$\alpha_k \sim N(0, 1.1), \text{ for } q \text{ k} = 1, \dots, K$$

Table C.2. Study 4B | Question Accuracy

Mixed Logistic Regression via lme4 (GLMER)

	odds ratios			(log odds)		
	Est.	2.5 %	97.5 %	Est.	2.5 %	97.5 %
(Intercept)	0.00 ***	0.00	0.01	-6.79 ***	-8.75	-4.83
MARK[cross]	4.71	0.73	30.46	1.55	-0.32	3.42
MARK[arrow]	31.95 ***	4.70	217.33	3.46 ***	1.55	5.38
SD (Intercept subject)	159.29			5.07		
SD (Intercept q)	3.02			1.10		
SD (Observations)	2.72			1.00		

Model ACCURACY ~ MARK + (1|subject) + (1|q)

n = 3913 R²(Conditional) = 0.9 R²(Marginal) = 0.06

* p < 0.05, ** p < 0.01, *** p < 0.001

C.3 Study 4C — | Supplementary Material

C.3.1 Accuracy Model

A Mixed Effects Logistic Regression Model indicates that shifting the y-axis of the TM graph from *orthogonal* to *triangular* significantly improves ACCURACY, while re-scaling the angles from *isosceles* to *equilateral* does not. (see Section 4.5.2.2).

$$\begin{aligned} \widehat{\text{accuracy}}_i &\sim \text{Binomial}(n = 1, \text{prob}_{\text{accuracy}=\text{correct}} = \hat{P}) \\ \log \left[\frac{\hat{P}}{1 - \hat{P}} \right] &= -7.44 \alpha_{j[i], k[i]} \\ \alpha_j &\sim N \left(-0.75 \gamma_1^\alpha (\text{SCALE}_{\text{equilateral}}) + 2.05 \gamma_2^\alpha (\text{SHAPE}_{\text{triangular}}), 6.57 \right) \\ &\text{, for subject } j = 1, \dots, J \\ \alpha_k &\sim N(0, 0.8), \text{ for } q, k = 1, \dots, K \end{aligned}$$

Table C.3. Study 4C | Question Accuracy

Mixed Logistic Regression via lme4 (GLMER)

	odds ratios			(log odds)		
	Est.	2.5 %	97.5 %	Est.	2.5 %	97.5 %
(Intercept)	0.00 ***	0.00	0.03	-7.44 ***	-11.34	-3.55
SCALE[equilateral]	0.47	0.07	3.03	-0.75	-2.61	1.11
SHAPE[triangular]	7.77 *	0.81	74.35	2.05 *	-0.21	4.31
SD (Intercept subject)	713.92			6.57		
SD (Intercept q)	2.22			0.80		
SD (Observations)	2.72			1.00		

Model ACCURACY \sim SHAPE * SCALE + (1|subject) + (1|q)

n = 3107 R²(Conditional) = 0.93 R²(Marginal) = 0.02

* p < 0.05, ** p < 0.01, *** p < 0.001

C.4 Study 4D — | Supplementary Material

C.4.1 Accuracy Model

A Mixed Effects Logistic Regression Model indicates that: (1) partially rotating a TM graph in space (e.g. by 45 degrees), and (2) changing the orientation of the y-axis from *orthogonal* to *triangular*, both significantly increase interpretation ACCURACY (see Section 4.6.2.2).

$$\widehat{\text{accuracy}}_i \sim \text{Binomial}(n = 1, \text{prob}_{\text{accuracy}=\text{correct}} = \hat{P})$$

$$\log \left[\frac{\hat{P}}{1 - \hat{P}} \right] = -4.05 \alpha_{j[i], k[i]}$$

$$\alpha_j \sim N \left(3.99 \gamma_1^\alpha (\text{ROTATE}_{45}) + 0.65 \gamma_2^\alpha (\text{ROTATE}_{90}) + 1.04 \gamma_3^\alpha (\text{SHAPE}_{\text{TRI}}), 2.96 \right)$$

, for subject j = 1, ..., J

$$\alpha_k \sim N(0, 0.65), \text{ for } q \text{ k} = 1, \dots, K$$

Table C.4. Study 4D | Question Accuracy

Mixed Logistic Regression via lme4 (GLMER)

	odds ratios			(log odds)		
	Est.	2.5 %	97.5 %	Est.	2.5 %	97.5 %
(Intercept)	0.02 ***	0.01	0.04	-4.05 ***	-4.90	-3.20
ROTATE[45]	54.31 ***	22.80	129.40	3.99 ***	3.13	4.86
ROTATE[90]	1.91	0.80	4.55	0.65	-0.22	1.51
SHAPE[triangular]	2.83 **	1.44	5.58	1.04 **	0.36	1.72
SD (Intercept subject)	19.36			2.96		
SD (Intercept q)	1.91			0.65		
SD (Observations)	2.72			1.00		

Model ACCURACY ~ ROTATE + SHAPE + (1|subject) + (1|q)

n = 5083 R²(Conditional) = 0.79 R²(Marginal) = 0.21

* p < 0.05, ** p < 0.01, *** p < 0.001

Appendix D

Multiple Choice Multiple Answer Scoring Strategy

The graph discovery (or graph comprehension) task used in Chapters 3 and 4 of this dissertation presents readers with a graph (the stimulus), a question, and a series of checkboxes. Participants are instructed to use the graph to answer the question, and respond by selecting all the checkboxes that apply, where each checkbox corresponds to a datapoint in the graph.

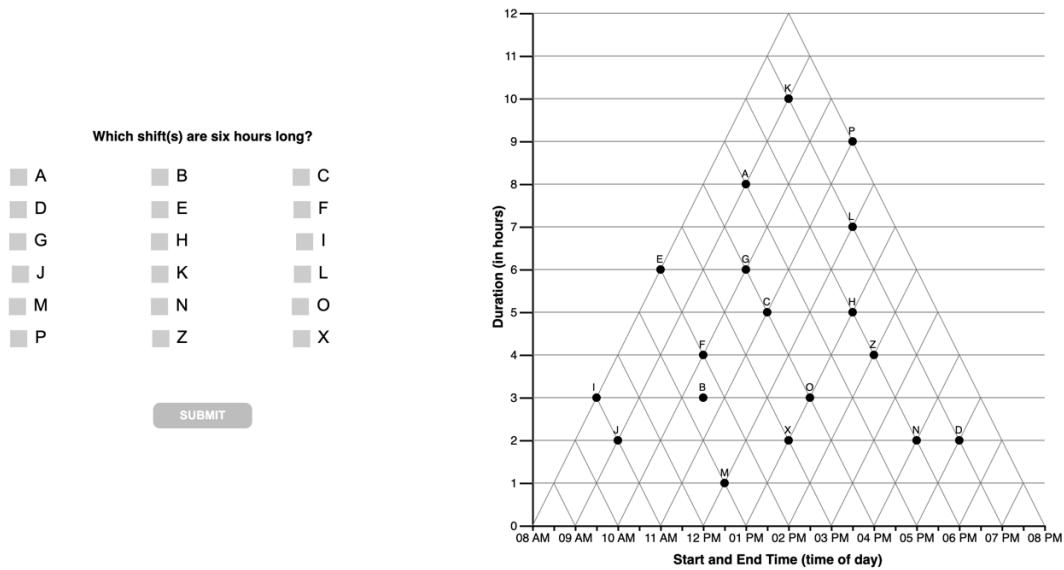


Figure D.1. Sample MCMA Question. Multiple Choice Multiple Answer questions allow respondents to construct an answer containing more information than the more limited response set of simple Multiple Choice questions.

In the psychometrics literature on tests and measures, the format of this type of question is referred to as Multiple Choice Multiple Answer (MCMA), (also: Multiple Response (MR) and Multiple Answer Multiple Choice (MAMC)). It has a number of properties that make it different from traditional Single Answer Multiple Choice (SAMC) questions, where the respondent marks a single response from a number of options. In particular, there are a number of very different ways that MCMA questions can be scored.

In traditional multiple choice (SAMC) questions, one point is given for selecting the option designated as correct, and zero points given for marking any of the alternative (i.e. distractor) options. Individual response options (*i*) on MCMA questions, however might be partially correct, while responses on other answer options within the same item might be incorrect. In MCMA, it is not obvious how to allocate points when the respondent marks a true-correct option (i.e. options that should be selected, denoted $\text{ } \checkmark$), as well as one or more false-correct options (i.e. options that should not be selected, denoted $\text{ } \times$). Should partial credit be awarded? If so, are options that respondents false-selected and false-unselected items equally penalized?

Schmidt and colleagueus (2021) performed a systematic literature review of publications proposing MCMA (or equivalent) scoring schemes, ultimately synthesizing over 80 sources into 27 distinct scoring approaches. Upon reviewing the benefits of trade-offs of each approach, for this study we choose to utilize two of the schemes: dichotomous scoring (Schmidt et. al. scheme #1), and partial scoring $[-1/q, + 1/p]$ (Schmidt et al. (2021) scheme #26) in order to measure both the strict correctness of an individual's response, as well as the most likely interpretation of the stimulus indicated by the unique pattern of their response.

D.1 Response Encoding

First, we note that the question type evaluated by Schmidt et al. (2021) is referred to as Multiple True-False (MTF), a variant of MCMA where respondents are presented with a question (stem) and series of response options with True/False (e.g. radio buttons) for each. Depending on

the implementation of the underlying instrument, it may or may not be possible for respondents to not respond to a particular option (i.e. leave the item ‘blank’). Although MTF questions have a different underlying implementation (and potentially different psychometric properties) they are identical in their mathematical properties; that is, responses to a MCMA question of ‘select all that apply’ can be coded as a series of T/F responses to each response option

Single Answer Multiple Choice (SAMC)	Multiple Answer Multiple Choice (MAMC)	Multiple True False (MTF)
Here is the question stem. <i>(Select one)</i>	Here is the question stem. <i>(Select all that apply)</i>	Here is the question stem. <i>(Select all that apply)</i>
<input type="radio"/> A	<input checked="" type="checkbox"/> A	A <input type="radio"/> True
<input type="radio"/> B	<input checked="" type="checkbox"/> B	B <input type="radio"/> True
<input type="radio"/> A,B,C,D	<input type="checkbox"/> C	C <input type="radio"/> True
<input checked="" type="radio"/> A and B only	<input type="checkbox"/> D	D <input type="radio"/> True
		<input type="radio"/> False
		<input checked="" type="radio"/> False
		<input type="radio"/> False
		<input checked="" type="radio"/> False

Figure D.2. MC vs MCMA vs MTF Question Formats.

In Figure D.2 we see an example of a question with four response options (A,B,C,D) in each question type. In the Multiple Choice (MC) approach (at left), there are four possible responses, given explicitly by the response options (respondent can select only one). With only four possible responses, we cannot entirely discriminate between all combinations of the underlying response variants we might be interested in, and must always choose an ‘ideal subset’ of possible distractors to present as response options. In the MCMA (middle) and MTF (at right), the same number of response options ($n=4$) yield more information about the respondent’s state of understanding. We can also see the equivalence between a MCMA and MTF format questions with the same response options. Options the respondent selects in MAMC are can be coded as T, and options they leave unselected can be coded as F. Thus, for response options (ABCD), a response of [AB] can also be encoded as [TTFF].

D.2 Scoring Schemes

In the sections that follow, we use the following terminology:

Properties of the Stimulus-Question

n number of response options (i.e. checkboxes)

p number of true-select options (i.e. should be selected)

q number of true-unselect options (i.e. should not be selected)

Properties of the Subject's Response

i number options in correct state ($0 \leq i \leq n$)

f resulting score

D.2.1 Dichotomous Scoring

Dichotomous Scoring is the strictest scoring scheme, where a response only receives points if it is exactly correct, meaning the respondent includes only correct-select options, and does not select any additional (i.e. incorrect-select) options that should not be selected. This is also known as all or nothing scoring, and importantly, it ignores any partial knowledge that a participant may be expressing through their choice of options. They may select some but not all of the correct-select options, and one or more but not all of the correct-unselect items, but receive the same score as a respondent selects none of the correct-select options, or all of the correct-unselect options. In this sense, dichotomous scoring tells us only about perfect knowledge, and ignores any indication of partial knowledge the respondent may be indicating through their selection of response options. In the studies described in this dissertation, we use the dichotomous scoring scheme to derive the ACCURACY score for each question, indicating if the participant's response was *precisely* correct, based on the triangular interpretation of the Triangular Model graph's coordinate system.

In Dichotomous Scoring:

- score for the question is either 0 or 1
- full credit is only given if all responses are correct; otherwise no credit
- does not account for partial knowledge. With increasing number of response options, scoring becomes stricter as each statement must be marked correctly

The algorithm for calculating a dichotomous score is given by:

$$f = \begin{cases} 1, & \text{if } i = n \\ 0, & \text{otherwise} \end{cases}$$

D.2.2 Partial Scoring [-1/n, +1/n]

Partial Scoring refers to a class of scoring schemes that award the respondent *partial credit* depending on the pattern of options they select. Schmidt et al. (2021) identify twenty-six different partial credit scoring schemes in the literature, varying in the range of possible scores, and the relative weighting of incorrectly selected (vs) incorrectly unselected options.

A particularly elegant approach to partial scoring is referred to as the approach **[-1/n, +1/n]** (Schmidt et al. (2021) #17). This approach is appealing in the context of this body of research, because it takes into account all information provided by the respondent: the pattern of what they select, as well as what they choose *not* to select.

In Partial scoring [-1/n, +1/n]

- Scores range from [-1, +1]
- One point is awarded if all options are correct
- One point is subtracted if all options are incorrect.

- Intermediate results are credited as fractions accordingly ($+1/n$ for each correct, $-1/n$ for each incorrect)
- This results in at chance performance (i.e. half of the given options marked correctly), being awarded 0 points

This scoring approach is more revealing given the motivating hypothesis that Triangular Graph readers start out with an incorrect (i.e. orthogonal, cartesian) interpretation of the coordinate system, and transition to a correct (i.e. triangular) interpretation. But the first step in making this transition is realizing the cartesian interpretation is incorrect, which may yield blank responses where the respondent is essentially saying, ‘there is no correct answer to this question’. Schmidt et al. (2021) describe this partial scoring scheme as the only scoring method (of the 27 described) where respondents’ scoring results can be interpreted as a percentage of their true knowledge.

One important drawback of this method is that a respondent may receive credit (a great deal of credit, depending on the number of answer options n) even if she did not select any options. In the case (such as ours) where there are *many more response options than there are options meant to be selected*, this partial scoring algorithm poses a challenge because the respondent can achieve an almost completely perfect score by selecting a small number of options that should not be selected.

The algorithm for calculating a partial scoring $[-1/n, +1/n]$ score is given by:

$$\begin{aligned} f &= (1/n * i) - (1/n * (n - i)) \\ &= (2i - n)/n \end{aligned}$$

D.2.3 Partial Scoring $[-1/q, +1/p]$

One drawback of the Partial Scoring $[-1/n, +1/n]$ approach is that treats the choice to select, and choice to not select options as equally indicative of the respondent’s understanding.

That is to say, incorrectly selecting one particular option is no more or less informative than incorrectly not-selecting a different item. This represents an important difference between MCMA (i.e. “select all correct options”) vs MTF (i.e. “Mark each option as true or false”) questions.

In this body of research, the selection of any particular option (remember options represent data points on the stimulus graph) is indicative of *a particular interpretation of the stimulus*. Incorrectly selecting an option indicates an interpretation of the graph with respect to that particular option. Alternatively, failing to select a correct option might mean the individual has a different interpretation, or that they failed to find all the data points consistent with the interpretation.

For this reason, we consider another alternative Partial Scoring scheme that takes into consideration only the selected statements, without penalizing statements incorrectly not selected. (See Schmidt et al. (2021) method #26; also referred to as the Morgan-Method) This partial scoring scheme takes into consideration that the most effort-free (or ‘default’) response for any given item is the null, or blank response. Blank responses indicate no understanding, perhaps confusion, or refusal to answer. These lack of responses are awarded zero credit. Whereas taking the action to select an incorrect option is effortful, and is indicative of incorrect understanding.

In Partial scoring [-1/q, +1/p]

- awards $+1/p$ points for each correctly selected option (p_s)
- subtracts $1/(n-p) = 1/q$ for each incorrectly selected option (q_s)
- only considers selected options; does not penalize nor reward unselected options

The algorithm for calculating a partial scoring $[-1/q, +1/p]$ score is given by:

$$f = (p_s/p) - (q_s/q)$$

The partial scoring $[-1/q, +1/p]$ scheme is the most appropriate partial scoring approach for determining an interpretation measure in this body of research because it allows us to to differentially penalize incorrectly selected and incorrectly not selected answer options, offering scores that maximally discriminate between alternative interpretations of the coordinate system.

D.3 Deriving the Interpretation Measure

In order to derive a single, categorical measure for a respondent's *interpretation* of the coordinate system of the stimuli in any particular question, we leverage a series of answer keys in combination with the partial scoring $[-1/q, +1/p]$ scoring scheme.

Based on the results of Study 1, we create answer keys for interpretation of the graph coordinate system we've previously observed, and thus expect we might observe in the future. These answer keys indicate the combination of response options that would be selected if the respondent was interpreting the stimulus in terms of each defined interpretation. For each question, we then use the partial scoring $[-1/q, +1/p]$ algorithm to calculate an interpretation-specific subscore based on the participant's response set. To decide on the (categorical) interpretation, we determine which interpretation subscore is the highest. This results in one and only one interpretation being assigned for each question. In cases where the subscores are non-discriminant (i.e. there is less than (0.5) points difference between the interpretation subscores), we assign an interpretation of *unknown*, indicating we are unable to determine how the participant interpreted the stimulus based on their chosen response set.

Bibliography

- Acarturk, C., Habel, C., & Cagiltay, K. (2008). Multimodal Comprehension of Graphics with Textual Annotations: The Role of Graphical Means Relating Annotations and Graph Lines. In G. Stapleton, J. Howse, & J. Lee (Eds.), *Diagrammatic Representation and Inference* (pp. 335–343). Springer Berlin Heidelberg. (Pages 34, 68).
- Acartürk, C. (2014). Towards a systematic understanding of graphical cues in communication through statistical graphs. *Journal of Visual Languages and Computing*, 25(2), 76–88. <https://doi.org/10.1016/j.jvlc.2013.11.006> (pages 53, 55, 68)
- Aigner, W., Miksch, S., Muller, W., Schumann, H., & Tominski, C. (2007). Visualizing time-oriented data. a systematic view. *Computers & Graphics*, 31(3), 401–409 (page 7).
- Allen, J. F. (1983). Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11), 823–832. <https://doi.org/http://doi.acm.org/10.1145/182.358434> (pages 44, 46)
- Anderson, J. R., & Bower, G. H. (1973). *Human Associative Memory*. V. H. Winston & Sons. (Page 25).
- Baird, J. C. (1970). *Psychophysical Analysis of Visual Space*. Pergamon Press. (Page 22).
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67, 1–48. <https://doi.org/10.18637/jss.v067.i01> (pages 95, 143)
- Bechtel, W., Abrahamsen, A., & Sheredos, B. (2018). Using Diagrams to Reason About Biological Mechanisms. In P. Chapman, G. Stapleton, A. Moktefi, S. Perez-Kriz, & F. Bellucci

- (Eds.), *Diagrammatic Representation and Inference* (pp. 264–279). Springer International Publishing. (Page 2).
- Beck, F., Burch, M., Diehl, S., & Weiskopf, D. (2016). A Taxonomy and Survey of Dynamic Graph Visualization. *Computer Graphics Forum*, 00(00), 1–27. <https://doi.org/10.1111/cgf.12791> (page 7)
- Bertin, J. (1967). *Sémiologie Graphique. Les diagrammes—les réseaux—les cartes*. Mouton. (Pages 17, 22, 33, 37, 134).
- Bertin, J. (1983). *Semiology of Graphics: Diagrams, Networks, Maps*. University of Wisconsin Press. (Pages 7, 17, 19, 22, 23, 25, 33, 37).
- Binder, K., Krauss, S., & Bruckmaier, G. (2015). Effects of visualizing statistical information – an empirical study on tree diagrams and 2×2 tables. *Frontiers in Psychology*, 6. <https://www.frontiersin.org/articles/10.3389/fpsyg.2015.01186> (page 2)
- Blackwell, A., & Engelhardt, Y. (2002). A Meta-Taxonomy for Diagram Research. *Diagrammatic Representation and Inference*, 47–64. http://link.springer.com/10.1007/978-1-4471-0109-3_3 (page 7)
- Blascheck, T., Kurzhals, K., Raschke, M., Burch, M., Weiskopf, D., & Ertl, T. (2017). Visualization of Eye Tracking Data: A Taxonomy and Survey. *Computer Graphics Forum*, 36(8), 260–284. <https://onlinelibrary.wiley.com/doi/full/10.1111/cgf.13079> (page 7)
- Boden, M. (2006). *Mind as machine: A history of cognitive science*. Oxford University Press. (Page 4).
- Bonin, S. (2000). Le développement de la graphique de 1967 à 1997. *Cybergeo : European Journal of Geography* (page 19).
- Boucheix, J.-M., & Schneider, E. (2009). Static and animated presentations in learning dynamic mechanical systems. *Learning and Instruction*, 19(2), 112–127 (page 34).
- Brehmer, M., & Munzner, T. (2013). A Multi-Level Typology of Abstract Visualization Tasks. *IEEE Transactions on Visualization and Computer Graphics*, 19(12) (page 37).

- Breslow, L. A., Trafton, J. G., & Ratwani, R. M. (2009). A perceptual process approach to selecting color scales for complex visualizations. *Journal of Experimental Psychology: Applied*, 15(1), 25–34 (page 34).
- Brinton, W. C. (1914). *Graphic methods for presenting facts*. The Engineering magazine company. (Page 15).
- Bürkner, P.-C. (2017). Brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, 80, 1–28. <https://doi.org/10.18637/jss.v080.i01> (page 97)
- Byrne, R. M. J., & Murray, M. A. (2005). Attention and Working Memory in Insight Problem-Solving. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 27(27). <https://escholarship.org/uc/item/0z89m2s8> (page 118)
- Carpenter, P. A., & Shah, P. (1998). A Model of the Cognitive and Perceptual Processes in Graph Comprehension. *Journal of Experimental Psychology: Applied*, 4(2), 75–100. <https://doi.org/10.1037/1076-898X.4.2.75> (pages 30, 35, 40, 43, 133)
- Castro, S. C., Hosseinpour, H., Quinan, P. S., & Padilla, L. (2021). Examining Effort in 1D Uncertainty Communication Using Individual Differences in Working Memory and NASA-TLX. *IEEE Transactions on Visualization and Computer Graphics*. <https://doi.org/10.1109/TVCG.2021.3114803> (page 122)
- Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, P. A. (1983). *Graphical Methods for Data Analysis*. Wadsworth Publ. Co. (Page 25).
- Chandler, D. (2017). *Semiotics: The basics* (Third edition). Routledge. (Pages 7, 8).
- Cheng, P. C.-H. (1996). Scientific Discovery With Law-Encoding Diagrams. *Creativity Research Journal*, 9(2-3), 145–162. <https://doi.org/10.1080/10400419.1996.9651169> (page 1)
- Cheng, P. C.-H. (2002). Electrifying diagrams for learning: Principles for complex representational systems. *Cognitive Science*, 26(6), 685–736. https://doi.org/10.1207/s15516709cog2606_1 (page 1)
- Cheng, P. C.-H. (2016). What Constitutes an Effective Representation? In M. Jamnik, Y. Uesaka, & S. Elzer Schwartz (Eds.), *Diagrammatic Representation and Inference* (pp. 17–31).

Springer International Publishing. https://doi.org/10.1007/978-3-319-42333-3_2.

(Page 1)

- Chi, E. (2000). A taxonomy of visualization techniques using the data state reference model. *IEEE Symposium on Information Visualization 2000. INFOVIS 2000. Proceedings*, 69–75 (page 7).
- Clark, A. (1997). *Being There: Putting Brain, Body and World Together Again*. MIT Press.
- (Page 10).
- Clark, A., & Chalmers, D. (1998). The Extended Mind. *Analysis*, 58(1), 7–19 (page 10).
- Cleveland, W. S., & McGill, R. (1984). Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *Journal of the American Statistical Association*, 79(387), 531–554. <https://doi.org/10.2307/2288400> (pages 21, 22, 33, 37, 39)
- Cleveland, W. S., & McGill, R. (1985). Graphical Perception and Graphical Methods for Analyzing Scientific Data. *Science*, 229(4716), 828–33 (page 22).
- Cleveland, W. S., & McGill, R. (1986). An Experiment In Graphical Perception. *International Journal of Man-Machine Studies*, 25(5), 491–500 (page 22).
- Cleveland, W. S., & McGill, R. (1987). Graphical Perception: The Visual Decoding of Quantitative Information on Graphical Displays of Data. *Journal of the Royal Statistical Society Series A-Statistics in Society*, 150, 192–229 (pages 22, 33, 37).
- Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, 12(5), 769–786. <https://doi.org/10.3758/BF03196772> (page 122)
- Coulson, S., & Cánovas, C. (2009). Understanding Timelines: Conceptual Metaphor and Conceptual Integration. *Cognitive Semiotics*, 5(1-2), 198–219. <http://www.degruyter.com/view/j/cogsem.2013.5.issue-1-2/cogsem.2013.5.12.198/cogsem.2013.5.12.198.xml> (page 45)

- Coutrot, A., Hsiao, J. H., & Chan, A. B. (2018). Scanpath modeling and classification with hidden Markov models. *Behavior Research Methods*, 50(1), 362–379 (page 43).
- Cox, D. R. (1978). Some Remarks on the Role in Statistics of Graphical Methods. *Applied Statistics*, 27(1), 4 (pages 15, 21).
- Croxton, F. E., & Stryker, R. E. (1927). Bar Charts versus Circle Diagrams. *Journal of the American Statistical Association*, 22(160), 473–482 (page 15).
- Curcio, F. R. (1987). Comprehension of mathematical relationships expressed in graphs. *Journal for Research in Mathematics Education*, 18(5), 382–393 (pages 35, 37, 41).
- de Leeuw, J. R. (2014). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, 1–12. <https://doi.org/10.3758/s13428-014-0458-y> (page 95)
- diSessa, A. (2004). Metarepresentation: Native competence and targets for instruction. *Cognition and Instruction*, (February 2014), 37–41. <https://doi.org/10.1207/s1532690xci2203> (page 172)
- diSessa, A., Hammer, D., & Sherin, B. (1991). Inventing graphing: Meta-representational expertise in children. *Journal of Mathematical Behavior*, 10, 117–160 (pages 41, 172).
- diSessa, A. A., & Sherin, B. L. (2000). Meta-representation: An introduction. *The Journal of Mathematical Behavior*, 19(4), 385–398. [https://doi.org/10.1016/S0732-3123\(01\)00051-7](https://doi.org/10.1016/S0732-3123(01)00051-7) (page 172)
- Duncker, K. (1945). "On problem solving". *Psychological Monographs*, 58:5 (Whole No. 270). (Page 85).
- Eells, W. C. (1926). The Relative Merits of Circles and Bars for Representing Component Parts. *Journal of the American Statistical Association*, 21(154), 119–132 (page 15).
- Engelhardt, Y. (2002). *The Language of Graphics* (Doctoral dissertation). University of Amsterdam. (Page 7).

- Eraslan, Yesilada, & Harper. (2016). Eye tracking scanpath analysis techniques on web pages: A survey, evaluation and comparison. *Journal of Eye Movement Research*, 9(1), 1–19.
<https://bop.unibe.ch/JEMR/article/view/2430> (page 43)
- Fauconnier, G., & Turner, M. (2002). *The way we think: Conceptual blending and the mind's hidden complexities*. Basic Books. (Page 137).
- Feeney, A., Hola, A. K. W., Liversedge, S. P., Findlay, J. M., & Metcalf, R. (2000). How people extract information from graphs: Evidence from a sentence-graph verification paradigm. In M. Anderson, P. C.-H. Cheng, & V. Haarslev (Eds.), *Theory and Application of Diagrams, Proceedings* (pp. 149–161). (Page 40).
- Follettie, J. (1986). Real-World Tasks of Statistical Graph-Using and Analytic Tasks of Graphics Research. *unpublished paper presented at the annual meeting of the National Computer Graphics Association, Anaheim, CA*. (pages 23, 37).
- Fox, A. R. (2020). A Psychology of Visualization or (External) Representation? *Proceedings of the Workshop on Visualization Psychology @ IEEE VIS*. <http://arxiv.org/abs/2009.13646> (page 37)
- Fox, A. R. (2023). Theories and Models of Graph Comprehension. In M. Chen, D. A. Szafir, R. Borgo, L. M. K. Padilla, D. J. Edwards, & B. Fisher (Eds.), *Visualization Psychology*. Springer. (Page xvi).
- Fox, A. R., de Vries, E., Lima, L., & Loker, S. (2016). Exploring Representations of Student Time-Use. In *Lecture Notes in Computer Science: Diagrammatic Reasoning and Inference*. (Page 45).
- Fox, A. R., & Hollan, J. D. (2018). Read it this way: Scaffolding comprehension for unconventional statistical graphs. In P. Chapman, G. Stapleton, A. Moktefi, S. Perez-Kriz, & F. Bellucci (Eds.), *Diagrammatic representation and inference* (pp. 441–457). Springer International Publishing. (Pages xvi, 34, 36, 82).

- Fox, A. R., & Hollan, J. D. (2023). Visualization Psychology: Foundations for an Interdisciplinary Research Programme. In M. Chen, D. A. Szafir, R. Borgo, L. M. K. Padilla, D. J. Edwards, & B. Fisher (Eds.), *Visualization Psychology*. Springer. (Page xvii).
- Fox, A. R., Hollan, J. D., & Walker, C. M. (2019). When graph comprehension is an insight problem. *Proceedings of the Annual Conference of the Cognitive Science Society* (pages xvi, 36, 131).
- Fox, A. R., & Van Den Berg, M. (2016). Representing Sequence: The Influence of Timeline Axis and Direction on Causal Reasoning in Litigation Law. *Proceedings of the Annual Meeting of the Cognitive Science Society* (page 45).
- Freedman, E. G., & Shah, P. (2002). Toward a model of knowledge-based graph comprehension. *Diagrammatic Representation and Inference*, 18–30. https://doi.org/10.1007/3-540-46037-3_3 (pages 30, 31)
- Friel, S. N., Curcio, F. R., & Bright, G. W. (2001). Making Sense of Graphs: Critical Factors Influencing Comprehension and Instructional Implications. *Journal for Research in Mathematics Education*, 32(2), 124–158 (pages 37, 38).
- Gillan, D. J., & Lewis, R. (1994). A Componential Model of Human Interaction with Graphs: 1. Linear Regression Modeling. *Human Factors*, 36(3), 419–440. <https://doi.org/10.1177/001872089403600303> (page 41)
- Glazer, N. (2011). Challenges with graph interpretation: A review of the literature. *Studies in Science Education*, 47(2), 183–210. <https://doi.org/10.1080/03057267.2011.605307> (page 38)
- Gooding, D. C. (2010). Visualizing Scientific Inference. *Topics in Cognitive Science*, 2(1), 15–35. <https://doi.org/10.1111/j.1756-8765.2009.01048.x> (page 2)
- Goodman, N. (1968). *Languages of Art: An Approach to a Theory of Symbols*. Hackett publishing. (Page 7).
- Harris, R. L. (1999). *Information Graphics: A Comprehensive Illustrated Reference*. Oxford University Press. (Page 7).

- Hegarty, M. (2011). The Cognitive Science of Visual-Spatial Displays: Implications for Design. *Topics in Cognitive Science*, 3(3), 446–474 (pages 38, 133).
- Hicks, K. L., Foster, J. L., & Engle, R. W. (2016). Measuring Working Memory Capacity on the Web with the Online Working Memory Lab (the OWL). *Journal of Applied Research in Memory and Cognition*, 5(4), 478–489. <https://doi.org/10.1016/j.jarmac.2016.07.010> (page 122)
- Hochpöchler, U., Schnitz, W., Rasch, T., Ullrich, M., Horz, H., McElvany, N., & Baumert, J. (2013). Dynamics of mental model construction from text and graphics. *European Journal of Psychology of Education*, 28(4), 1105–1126. <https://doi.org/10.1007/s10212-012-0156-z> (page 41)
- Hollan, J. D., Hutchins, E., & Kirsh, D. (2000). Distributed cognition: Toward a new foundation for human-computer interaction research. *Transactions on Computer-Human Interaction (TOCHI)*, 7(2), 174–196 (page 11).
- Holmqvist, K., & Andersson, R. (2017). Introduction. In *Eye-tracking: A comprehensive guide to methods, paradigms and measures*. Lund Eye-Tracking Research Institute. (Page 42).
- Hoopes, J. (1991). *Peirce on Signs: Writings on Semiotic by Charles Sanders Peirce*. The University of North Carolina Press. (Pages 9, 10).
- Hsieh, H.-F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative health research*, 15(9), 1277–88. <https://doi.org/10.1177/1049732305276687> (page 78)
- Hullman, J., Adar, E., & Shah, P. (2011). The Impact of Social Information on Visual Judgments. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1461–1470 (page 36).
- Hullman, J., Kay, M., Kim, Y., & Shrestha, S. (2018). Imagining Replications: Graphical Prediction & Discrete Visualizations Improve Recall & Estimation of Effect Uncertainty. *IEEE Transactions on Visualization and Computer Graphics*, 24(1), 446–456 (page 36).

- Hutchins, E. (2001). Cognition, Distributed. In *International Encyclopedia of the Social & Behavioral Sciences* (pp. 2068–2072). Elsevier. (Page 10).
- Hutchins, E. (1995). Cognition in the Wild. *MIT Press*, 1–5. <https://doi.org/10.1023/A:1008642111457> (pages 2, 10)
- Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2(3), 194–203. <https://doi.org/10.1038/35058500> (page 34)
- Jeffreys, t. l. H. (1961). *Theory of Probability* (3rd edition). Oxford University Press. (Page 98).
- Joint Committee on Standards for Graphic Presentation. (1915). *Publications of the American Statistical Association*, 14(112), 790–797. <https://doi.org/10.2307/2965153> (page 15)
- Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, 99(1), 122–149. <https://doi.org/10.1037/0033-295X.99.1.122> (page 117)
- Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4), 329–354. <https://doi.org/http://dx.doi.org/10.1037/0033-295X.87.4.329> (page 42)
- Kaiser, D. (2005). Physics and Feynman's diagrams. *American Scientist*, 93(2), 156–165. <https://doi.org/10.1511/2005.2.156> (page 2)
- Kim, Y.-S., Walls, L. A., Krafft, P., & Hullman, J. (2019). A Bayesian Cognition Approach to Improve Data Visualization. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14 (page 36).
- Kintsch, W. (1998). Modeling comprehension processes: The construction-integration model'. In *Comprehension. A paradigm for cognition*. (pp. 93–120). Cambridge University Press. (Page 30).
- Kirsh, D. (1990). When is Information Explicitly Represented? In *The Vancouver Studies in Cognitive Science* (pp. 340–365). Oxford University Press. (Page 6).
- Kirsh, D. (2006). Implicit and Explicit Representation. *Encyclopedia of Cognitive Science*, 478–481 (page 6).

- Kirsh, D. (2010). Thinking with external representations. *AI and Society*, 25(4), 441–454. <https://doi.org/10.1007/s00146-010-0272-8> (page 1)
- Kong, N., & Agrawala, M. (2012). Graphical overlays: Using layered elements to aid chart reading. *IEEE Transactions on Visualization and Computer Graphics*, 18(12), 2631–2638. <https://doi.org/10.1109/TVCG.2012.229> (pages 34, 53, 68)
- Kosslyn, S. M. (1985). Graphics and Human Information Processing: A Review of Five Books. *Journal of the American Statistical Association*, 25(4), R134–R136 (page 24).
- Kosslyn, S. M. (1989). Understanding charts and graphs. *Applied Cognitive Psychology*, 3(3), 185–225. <https://doi.org/10.1002/acp.2350030302> (pages 25, 132, 136)
- Kruskal, W. (1975). Visions of Maps and Graph. *Auto-Carto II, Proceedings of the International Symposium on Computer Assisted Cartography*, ed. J. Kavaliunas, Washington, D.C.: U.S. Bureau of the Census and American Congress on Survey and Mapping., 27–36 (pages 15, 21).
- Kulpa, Z. (1997). Diagrammatic Representation for a Space of Intervals. *Machine Graphics & Vision*, 5–24 (pages 47, 50, 152).
- Kulpa, Z. (2000). A Diagrammatic Notation for Interval Algebra. In M. Anderson, P. Cheng, & V. Haarslev (Eds.), *Theory and Application of Diagrams* (pp. 471–474). Springer Berlin Heidelberg. (Page 151).
- Kulpa, Z. (2001). Diagrammatic representation for interval arithmetic. *Linear Algebra and Its Applications*, 324, 55–80 (pages 50, 152).
- Kulpa, Z. (2006). A diagrammatic approach to investigate interval relations. *Journal of Visual Languages and Computing*, 17(5), 466–502. <https://doi.org/10.1016/j.jvlc.2005.10.004> (page 47)
- Lakoff, G., & Johnson, M. (1980). *Metaphors We Live By*. University of Chicago Press. <https://doi.org/978-0226468013>. (Page 137)

- Larkin, J., & Simon, H. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, 99, 65–99. <http://onlinelibrary.wiley.com/doi/10.1111/j.1551-6708.1987.tb00863.x/abstract> (pages 1, 3, 6, 16, 47)
- Latour, B., & Woolgar, S. (1987). *Laboratory Life. The Construction of Scientific Facts*. <https://press.princeton.edu/books/paperback/9780691028323/laboratory-life>. (Page xiii)
- Lindsay, P. H., & Norman, D. (1972). *Human information processing: An introduction to psychology*. Academic Press. (Pages 4, 25).
- Liu, Z., Nersessian, N., & Stasko, J. (2007). Distributed cognition as a theoretical framework for information visualization. *IEEE transactions on visualization and computer graphics*, 14(6), 1173–80 (page 11).
- Livingston, M. A., Matzen, L. E., Harrison, A., Lulushi, A., Daniel, M., Dass, M., Brock, D., & Decker, J. W. (2020). A Study of Perceptual and Cognitive Models Applied to Prediction of Eye Gaze within Statistical Graphs. *ACM Symposium on Applied Perception 2020*, 1–9. <https://doi.org/10.1145/3385955.3407931> (page 34)
- Lohse, G. (1993). A cognitive model for understanding graphical perception. *Human-Computer Interaction*, 8(4), 353–388. https://doi.org/10.1207/s15327051hci0804_3 (pages 41, 133, 136)
- Lohse, G. (1997). Models of Graphical Perception. In *Handbook of Human-Computer Interaction* (Second Edition). Elsevier Science & Technology. (Page 136).
- Lohse, G. L. (1997). The role of working memory on graphical information processing. *Behaviour & Information Technology*, 16(6), 297–308. <https://doi.org/10.1080/014492997119707> (pages 117, 128, 129)
- Ludewig, U. (2018). *Understanding Graphs: Modeling Processes, Prerequisites and Influencing Factors of Graphicacy* (Doctoral dissertation). Universität Tübingen. Tübingen, Germany. <https://publikationen.uni-tuebingen.de/xmlui/handle/10900/84624>. (Page 35)
- Mackinlay, J. (1986). Automating the Design of Graphical Presentations of Relational Information. *ACM Transactions on Graphics*, 5, 110–141 (page 33).

- Makowski, D., Ben-Shachar, M. S., Chen, S. H. A., & Lüdecke, D. (2019). Indices of Effect Existence and Significance in the Bayesian Framework. *Frontiers in Psychology*, 10. <https://www.frontiersin.org/articles/10.3389/fpsyg.2019.02767> (page 98)
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. The MIT Press. <http://mitpress.universitypressscholarship.com/view/10.7551/mitpress/9780262514620.001.0001/upso-9780262514620>. (Pages 16, 25, 28)
- Massironi, M. (2001). *The Psychology of Graphic Images*. Psychology Press. (Page 7).
- Mautone, P. D., & Mayer, R. E. (2007). Cognitive aids for guiding graph comprehension. *Journal of Educational Psychology*, 99(3), 640–652. <https://doi.org/10.1037/0022-0663.99.3.640> (pages 34, 40, 53–55, 68)
- Merwin, D. H., & Wickens, C. D. (1993). Comparison of Eight Color and Gray Scales for Displaying Continuous 2D Data. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 37(19), 1330–1334 (page 34).
- Minsky, M. (1974). A Framework for Representing Knowledge. In *P. H. Winston (Ed.). The psychology of computer vision*. McGraw-Hill. <https://dspace.mit.edu/handle/1721.1/6089>. (Page 135)
- Nauta, D. (1972). *The Meaning of Information*. Mouton. (Page 12).
- Neisser, U. (1967). *Cognitive Psychology*. Appleton-Century-Crofts. (Page 4).
- Norman, D. A. (1993). *Things That Make Us Smart: Defending Human Attributes in the Age of the Machine*. Addison-Wesley Longman Publishing Co., Inc. (Pages 2, 4).
- Norman, D. A., & Rumelhart, D. E. (1975). *Explorations in Cognition*. W. H. Freeman & Company. (Page 135).
- Núñez & Cooperrider. (2013). The tangle of space and time in human cognition. *Trends in Cognitive Sciences*, 17(5), 220–9. <https://doi.org/10.1016/j.tics.2013.03.008> (page 45)

- Ohlsson, S. (1992). Information-processing explanations of insight and related phenomena. In *Advances in the Psychology of Thinking* (pp. 1–44). <https://doi.org/NEED>. (Pages 84, 85, 105)
- Oswald, F. L., McAbee, S. T., Redick, T. S., & Hambrick, D. Z. (2015). The development of a short domain-general measure of working memory capacity. *Behavior Research Methods*, 47(4), 1343–1355. <https://doi.org/10.3758/s13428-014-0543-2> (page 121)
- Padilla, L., Castro, S. C., & Hosseinpour, H. (2021). A review of uncertainty visualization errors: Working memory as an explanatory theory. In *Psychology of Learning and Motivation* (pp. 275–315). Elsevier. <https://doi.org/10.1016/bs.plm.2021.03.001>. (Page 173)
- Padilla, L. M., Creem-Regehr, S. H., Hegarty, M., & Stefanucci, J. K. (2018). Decision making with visualizations: A cognitive framework across disciplines. *Cognitive Research: Principles and Implications*, 3(1), 29 (page 173).
- Palmer, S. E. (1978). Fundamental Aspects of Cognitive Representation. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and Categorization* (pp. 259–302). Erlbaum. (Pages 1, 6, 47, 134).
- Palmer, S. E., & Kimchi, R. (1986). The Information Processing Approach to Cognition. In *Approaches to Cognition: Contrasts and Controversies* (p. 38). Lawrence Erlbaum Associates. (Page 4).
- Palsky, G. (2019). Jacques Bertin, from classical training to systematic thinking of graphic signs. *Cartography and Geographic Information Science*, 46(2), 189–193 (page 17).
- Parsons, P., & Sedig, K. (2014a). Common Visualizations: Their Cognitive Utility. In *Handbook of Human Centric Visualization* (pp. 671–691). (Page 7).
- Parsons, P., & Sedig, K. (2014b). Distribution of Information Processing While Performing Complex Cognitive Activities with Visualization Tools. In *Handbook of Human Centric Visualization* (pp. 671–691). (Page 14).

- Peebles, D., & Cheng, P. C.-H. (2002). Extending task analytic models of graph-based reasoning: A cognitive model of problem solving with Cartesian graphs in ACT-R/PM. *Cognitive Systems Research*, 3(1), 77–86 (page 133).
- Peebles, D., & Cheng, P. C.-H. (2003). Modeling the effect of task and graphical representation on response latency in a graph reading task. *Human Factors*, 45(1), 28–46. <https://doi.org/10.1518/hfes.45.1.28.27225> (pages 42, 43, 136)
- Pike, W. A., Stasko, J., Chang, R., & O'Connell, T. A. (2009). The Science of Interaction. *Information Visualization*, 8(4), 263–274 (page 34).
- Pinker, S. (1990). Theory of Graph Comprehension. In R. Freedle (Ed.), *Artificial Intelligence and the Future of Testing* (pp. 73–126). Erlbaum. (Pages 7, 26, 28, 29, 32, 56, 57, 82, 132–136).
- Postigo, Y., & Pozo, J. I. (2004). On the Road to Graphicacy: The Learning of Graphical Representation Systems. *Educational Psychology*, 24(5), 623–644 (page 36).
- Qiang, Y., Delafontaine, M., Versichele, M., De Maeyer, P., & Van de Weghe, N. (2012). Interactive Analysis of Time Intervals in a Two-Dimensional Space. *Information Visualization*, 11(4), 255–272. <https://doi.org/10.1177/1473871612436775> (pages 45, 47, 70, 75, 80, 86, 151)
- Qiang, Y., Valcke, M., De Maeyer, P., & Van de Weghe, N. (2014). Representing time intervals in a two-dimensional space: An empirical study. *Journal of Visual Languages & Computing*, 25(4), 466–480. <https://doi.org/10.1016/j.jvlc.2014.01.001> (pages 47, 48)
- R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. <https://www.R-project.org/>. (Pages 95, 143)
- Ratwani, R. M., & Trafton, J. G. (2008). Shedding light on the graph schema: Perceptual features versus invariant structure. *Psychonomic Bulletin & Review*, 15(4), 757–62. <https://doi.org/10.3758/pbr.15.4.757> (pages 135, 136, 173)
- Ratwani, R. M., Trafton, J. G., & Boehm-Davis, D. A. (2008). Thinking Graphically: Connecting Vision and Cognition during Graph Comprehension. *Journal of Experimental Psychology*:

- Applied*, 14(1), 36–49. <https://doi.org/10.1037/1076-898X.14.1.36> (pages 40, 41, 43, 136)
- Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3), 372–422 (page 42).
- Resnick, I., Newcombe, N. S., & Shipley, T. F. (2016). Dealing with Big Numbers: Representation and Understanding of Magnitudes Outside of Human Experience. *Cognitive Science*, 1–22. <https://doi.org/10.1111/cogs.12388> (page 2)
- Rogers, Y. (2008). When the external entered HCI: Designing effective representations. In T. Erickson & D. W. McDonald (Eds.), *HCI Remixed: Reflections on Works That Have Influenced the HCI Community* (pp. 275–280). MIT Press. (Page 12).
- Roth, W.-M. (2003). Toward an Anthropology of Graphing. In W.-M. Roth (Ed.), *Toward an Anthropology of Graphing: Semiotic and Activity-Theoretic Perspectives* (pp. 1–21). Springer Netherlands. (Pages 2, 39, 52).
- Roth, W.-M., & Bowen, G. M. (2003). When Are Graphs Worth Ten Thousand Words? An Expert-Expert Study. *Cognition and Instruction*, 21(4), 429–473. https://doi.org/10.1207/s1532690xci2104_3 (page 39)
- Scaife, M., & Rogers, Y. (1996). External cognition: How do graphical representations work? *International Journal of Human-Computer Studies*, 185–213. <http://www.sciencedirect.com/science/article/pii/S1071581996900488> (pages 1, 6)
- Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals and understanding: An inquiry into human knowledge structures*. Lawrence Erlbaum. (Page 135).
- Scheiter, K., & Eitel, A. (2017). The Use of Eye Tracking as a Research and Instructional Tool in Multimedia Learning. <https://doi.org/10.4018/978-1-5225-1005-5.ch008> (page 42)
- Schmidt, D., Raupach, T., Wiegand, A., Herrmann, M., & Kanzow, P. (2021). Relation between examinees' true knowledge and examination scores: Systematic review and exemplary calculations on Multiple-True-False items. *Educational Research Review*, 34, 100409. <https://doi.org/10.1016/j.edurev.2021.100409> (pages 93, 199, 202)

- Sedig, K., & Parsons, P. (2013). Interaction Design for Complex Cognitive Activities with Visual Representations: A Pattern-Based Approach. *AIS Transactions on Human-Computer Interaction*, 5(2), 84–133 (pages 13, 14, 34).
- Sedig, K., & Parsons, P. (2016). *Design of Visualizations for Human-Information Interaction: A Pattern-Based Framework*. Morgan & Claypool Publishers. (Pages 13, 14).
- Sedig, K., Parsons, P., & Babanski, A. (2012). Towards a Characterization of Interactivity in Visual Analytics. *Journal of Multimedia Processing and Technologies, Special Issue on Theory and Application of Visual Analytics*, 3(1), 12–28 (page 13).
- Sedig, K., Parsons, P., Dittmer, M., & Haworth, R. (2014). Human-Centered Interactivity of Visualization Tools: Micro and Macro-level Considerations. In *Handbook of Human Centric Visualization* (pp. 671–691). (Page 14).
- Shah, P. (1997). A Model of the Cognitive and Perceptual Processes in Graphical Display Comprehension. In *Reasoning with diagrammatic representations II* (pp. 94–101). AAI Press. (Page 30).
- Shah, P. (2002). Graph Comprehension: The Role of Format, Content and Individual Differences. In M. Anderson, B. Meyer, & P. Olivier (Eds.), *Diagrammatic Representation and Reasoning* (pp. 173–185). Springer London. http://link.springer.com/10.1007/978-1-4471-0109-3_10. (Pages 30, 31, 34, 36)
- Shah, P., & Carpenter, P. A. (1995). Conceptual Limitations in Comprehending Line Graphs. *Journal of Experimental Psychology: General*, 124(1), 43–61. <https://doi.org/10.1037/0096-3445.124.1.43> (page 39)
- Shah, P., & Freedman, E. G. (2011). Bar and line graph comprehension: An interaction of top-down and bottom-up processes. *Topics in Cognitive Science*, 3(3), 560–578 (pages 133, 136).
- Shah, P., Freedman, E. G., & Vekiri, I. (2005). The Comprehension of Quantitative Information in Graphical Displays. In P. Shah (Ed.), *The Cambridge Handbook of Visuospatial*

- Thinking* (426–476, Chapter xviii, 561 Pages). Cambridge University Press, New York, NY. (Pages 32, 56).
- Shah, P., & Hoeffner, J. (2002). Review of graph comprehension research: Implications for instruction. *Educational Psychology Review*, 14(1), 47–69 (pages 2, 38, 52, 56).
- Shneiderman, B. (1996). The Eyes Have It: A Task By Data Type Taxonomy for Information Visualization. *Proceedings 1996 IEEE Symposium on Visual Languages*, 336–343 (page 7).
- Simkin, D., & Hastie, R. (1987). An Information-Processing Analysis of Graph Perception. *Journal of the American Statistical Association*, 82(398), 454–465. <https://doi.org/10.1080/01621459.1987.10478448> (pages 23, 37, 39)
- Sims, V. K., & Hegarty, M. (1997). Mental animation in the visuospatial sketchpad: Evidence from dual-task studies. *Memory & Cognition*, 25(3), 321–332. <https://doi.org/10.3758/bf03211288> (page 117)
- Smart, S., & Szafir, D. A. (2019). Measuring the Separability of Shape, Size, and Color in Scatterplots. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14 (page 34).
- Stevens, S. S. (1975). *Psychophysics*. John Wiley. (Page 22).
- Strobel, B., Lindner, M. A., Saß, S., & Köller, O. (2018). Task-irrelevant data impair processing of graph reading tasks: An eye tracking study. *Learning and Instruction*, 55, 139–147. <https://doi.org/10.1016/j.learninstruc.2017.10.003> (page 42)
- Strobel, B., Sass, S., Lindner, M. A., & Koeller, O. (2016). Do Graph Readers Prefer the Graph Type Most Suited to a Given Task? Insights from Eye Tracking. *Journal of Eye Movement Research*, 9(4), 4. <https://doi.org/10.16910/jemr.9.4.4> (page 40)
- Stull, A. T., Barrett, T., & Hegarty, M. (2013). Usability of concrete and virtual models in chemistry instruction. *Computers in Human Behavior*, 29(6), 2546–2556. <https://doi.org/10.1016/j.chb.2013.06.012> (page 1)

- Szafir, D. A., Haroz, S., Gleicher, M., & Franconeri, S. (2016). Four types of ensemble coding in data visualizations. *Journal of Vision*, 16(5), 11–11. <https://doi.org/10.1167/16.5.11> (pages 34, 37)
- Tory, M., & Moller, T. (2004). Rethinking Visualization: A High-Level Taxonomy. *IEEE Symposium on Information Visualization*, 151–158 (page 7).
- Tufte, E. (1983). *Visual Display of Quantitative Information* (First). Graphics Paper Press LLC. (Pages 25, 138).
- Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley. (Page 15).
- Tversky, B. (2011). Visualizing Thought. *Topics in Cognitive Science*, 3(3), 499–535 (page 167).
- Tversky, B., Morrison, J. B., & Betrancourt, M. (2002). Animation: Can it facilitate? *International Journal of Human-Computer Studies*, 57(4), 247–262 (page 34).
- Ullman, S. (1984). Visual routines. *Cognition*, 18(1), 97–159 (pages 16, 135).
- Van de Weghe, N., Docter, R., De Maeyer, P., Bechtold, B., & Ryckbosch, K. (2007). The triangular model as an instrument for visualising and analysing residuality. *Journal of Archaeological Science*, 34(4), 649–655. <https://doi.org/10.1016/j.jas.2006.07.007> (pages 47, 151, 152)
- Velez, M.C., Silver, D., & Tremaine, M. (2005). Understanding visualization through spatial ability differences. *VIS 05. IEEE Visualization*, 2005., 511–518 (page 35).
- von Huhn, R. (1927). Further Studies in the Graphic Use of Circles and Bars: I: A Discussion of the Eells' Experiment. *Journal of the American Statistical Association*, 22(157), 31–36 (page 15).
- Wainer, H., & Thissen, D. (1981). Graphical Data Analysis. *Annual Review of Psychology*, 32(1), 191–241 (page 24).
- Wainer, H. (1992). Understanding Graphs and Tables. *Educational Researcher*, 21(1), 14–23. <https://doi.org/10.3102/0013189X021001014> (pages 37, 41)

- Washburne, J. N. (1927). An experimental study of various graphic, tabular, and textual methods of presenting quantitative material. *Journal of Educational Psychology*, 18(6), 361–376 (page 15).
- Wright, J. C., & Murphy, G. L. (1984). The Utility of Theories in Intuitive Statistics: The Robustness of Theory-Based Judgments. *Journal of Experimental Psychology: General*, 113(2), 301–322 (page 36).
- Zacks, J., & Tversky, B. (1999). Bars and lines: A study of graphic communication. *Memory & Cognition*, 27(6), 1073–9. <https://doi.org/10.3758/BF03201236> (page 40)
- Zhang, J., & Norman, D. (1994). Representations in Distributed Cognitive Tasks. *Cognitive Science*, 18(1), 87–122 (page 6).