UNIVERSITY OF CALIFORNIA SAN DIEGO

Adaptive Visualization Strategies across Drawings, Diagrams, and Data Visualizations

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

in

Experimental Psychology

by

Holly Huey

Committee in charge:

Professor Judith E. Fan, Co-Chair
Professor Caren M. Walker, Co-Chair
Professor Timothy F. Brady
Professor Adena Schachner
Professor Hiajun Xia

2024

The Dissertation of Holly Huey is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2024

DEDICATION

To my moms, for always encouraging and supporting me no matter how meandering my life choices have been and continue to be.

And to my younger sister and brother, for always keeping me humble:

“You get a *hood* for your *Master*’s Degree? That’s sketchy...”

“You could just buy yourself Renaissance robes instead of getting your PhD. All academics seem to want to play dress up for Ren Faire.”

EPIGRAPH

“Styles, like languages, differ in the sequence of articulation and in the number of questions they allow the artist to ask; and so complex is the information that reaches us from the visible world that no picture will ever embody it all. This is not due to the subjectivity of vision but to its richness.”

Ernest Gombrich (1960)

“By stripping down an image to its essential ‘meaning’, an artist can amplify that meaning in a way that realistic art can’t.”

Scott McCloud (1993)

“That piece of paper becomes *our* product, not yours and not mine, so it’s a genuine collaboration that we’re both invested in.”

Barbara Tversky (2019)

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

It is incredible to look back on the last five years of the PhD. I have had so many brilliant mentors, both formally and informally, but Judy Fan above all has nothing short of changed my life. During the first of our many walks, she patiently listened to me ramble about how I had struggled to explain to others how my prior art apprenticeships and *very, very* scattered research experiences at that point somehow felt linked by my singular fascination about how people discover and communicate ideas through pictures. While I'm certain most of that came out as quarter-baked nonsense, Judy—since that moment and to today—gave me the confidence to hold on to this gut feeling as an asset to explore through different scientific domains, methods, and techniques. I am continually inspired by her insistence to do science *right*—from her careful crafting of scientific narratives in manuscripts and slide-smithing to her tireless dedication to her students. Thank you for teaching me that no matter the rejections, that if we did the science *right* we could always be confident that our research would find a good home. Thank you for teaching me to celebrate the blood (not literally), sweat, and tears (definitely some) of a finished manuscript more than the semi-arbitrary acceptance of a conference, journal, or grant. For your endless encouragements even through your truly superhuman late night pushes, I want to extend my most heartfelt thank you. I'm really looking forward to our final YERM (FERM?) but also having the opportunity to catch up every once in a while in the Bay Area.

I wish to express my deepest gratitude to Caren Walker for adopting me into her lab twice and for braving Overleaf for me. I have grown so much from her generosity as a scientist—from her expansive knowledge of an astonishing number of literature domains to connecting me with her astonishing number of collaborators—and from her generosity as a person. One of my favorite grad school core memories is telling her at the Society for Research in Child Development that I was accepting UCSD's offer and how she pulled me into a HUGE hug and then very excitedly introduced me to everyone

at the conference in the immediate vicinity. I would also like to thank my committee members—Tim Brady, Adena Schachner, and Haijun Xia—who have offered insightful feedback and encouragement throughout the years but also, due to world circumstances, have been flexible with me delivering every PhD milestone to date via Zoom. I am honored to present my dissertation defense to you in-person as my final milestone.

Thank you to my incredible, long-time collaborators, who have helped shaped both my research and who I am as a scientist. To Justin Yang, Xuanchen Lu, Kushin Mukherjee, Hannah Lloyd, and Lauren Oey—thank you for engaging in deep scientific conversation over many coffees, hikes, Zooms, and long campus walks. Science has been more fun with you. I have also been lucky to have a constellation of mentors throughout my research career, like Bria Long, Julian Jara-Ettinger, Jonathan Kominsky, Matthew Fisher, Mackenzie Leake, and Anh Truong, whom I have retrospectively realized catalyzed major inflection points in my life trajectory. Additionally, I have no words to truly express how lucky and honored I feel to have been part of the first generation of scientists in the Cognitive Tools Lab: Will McCarthy, Haoliang Wang (thank you for helping me fix my Overleaf compiling errors in the eleventh hour!), Felix Binder, Hannah Lloyd, Sebastian Holt, Justin Yang, Erik Brockbank, Zoe Tait, Xuanchen Lu, Arnav Verma, and Lauren Oey. Thank you for being such kind friends but also mentors who have taught me more than I can describe. Thank you also to all the undergrads that I have collaborated with over the years, who have shaped who I am and want to continue growing into as a mentor—in particular, thank you to Rio Aguina-Kang, Vivian Leung, Marcus Franco, Sara Okun, Anaïs Kessler, and Olivia Miller who witnessed me explore a number of mentoring techniques and offered honest and insightful feedback when I asked.

There is a non-zero chance that I would not be alive or unmaimed if not for the support and wilderness teachings of my hiking friends like Lauren Oey, Philip Belzeski, Leo Kleiman-Lynch, Emily Laino, Hannah Lloyd, Erik Brockbank, Alex Carstensen, and Janna Dickenson—thank you for letting me borrow so much gear over the years! In my wildest dreams, I had not imagined being able to embark on adventures like the John Muir Trail, Grand Canyon Rim-to-Rim, Half Dome, or Trans-Catalina Trail, and I have loved that I share many of these memories with you (Ren, in particular, my backpacking buddy who offloads extra weight into my backpack whenever they can if I am being "too cheerful"). So many thanks also to my cohort mates (and Alex Rett), my friends from Whidbey Island to New York City (thank you, Rosie and Katja, for keeping me sane this summer), and Molly, Maya, & Zuko for the essential emotional support.

An enormous thank you to Manasvi Sridhar, for their billowing but tender love and for sharing their voracious adoration for all the tiny beautiful things in the world—I would not have known how expansive life can be, both externally and internally, without you. And to Kody Kodkany, for his endless patience with my latest hiking obsessions, for his sage wisdom including calling me out when I am "stressing out about the wrong things", and for making me a coffee addict.

And most importantly, thank you to my family. I would not have taken the chances that I have in life without knowing that it has always been okay for me to come home.

Chapter 1, in full, is currently being prepared for submission for publication of the material. An earlier version of the project was published as Yang, J.& Fan, J. E.. (2021). Visual communication of object concepts at different levels of abstraction. *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society.* Cognitive Science Society. The dissertation author was the primary investigator and author of this updated material.

Chapter 2, in full, is a reprint of material as it appears in Huey, H., Lu, X., Walker, C.M., & Fan, J.E. (2023). Explanatory drawings prioritize functional properties at the expense of visual fidelity. *Cognition.* An earlier version of the project was published as Huey, H., Walker, C., & Fan, J. (2021). How do the semantic properties of visual explanations guide causal inference? *Proceedings of the 43rd Annual Conference of the Cognitive Science Society.* Cognitive Science Society. The dissertation author was the primary investigator and author of this material.

Chapter 3, in full, is currently being prepared for submission for publication of the material. An earlier version of the project was published as Huey*, H., Oey*, L., Lloyd, H., & Fan, J. (2023). How do communicative goals guide which data visualizations people think are effective? *Proceedings of the 45th Annual Meeting of the Cognitive Science Society*. Cognitive Science Society. The dissertation author was the primary investigator and author of this material.

VITA

2012	Bachelor of Arts, Liberal Arts, St. John's College
2021	Master of Arts, Experimental Psychology, University of California San Diego
2024	Doctor of Philosophy, Experimental Psychology, University of California San Diego

PUBLICATIONS

Huey, H., Oey, L. A., Lloyd, H. S., & Fan, J. E. (n.d). Evaluating communicative constraints on data visualization design. [in preparation].

*Huey, H., *Yang, J., Lu, X., & Fan, J. E. (n.d). Visual communication of object concepts at different levels of abstraction. [in preparation].

Huey, H., Leake, M., Aneja, D., Fisher, M. D., & Fan, J. E. (2024). How do video content creation goals impact which concepts people prioritize for generating b-roll imagery? *Proceedings of the 16th Conference on Creativity & Cognition*, 542–549.

Long, B., Fan, J. E., Huey, H., Chai, Z., & Frank, M. C. (2024). Parallel developmental changes in children's production and recognition of line drawings of visual concepts. *Nature Communications*, *15*(1), 1191.

*Mukherjee, K., *Huey, H., *Lu, X., Vinker, Y., Aguina-Kang, R., Shamir, A., & Fan, J. (2024). Seva: Leveraging sketches to evaluate alignment between human and machine visual abstraction. *Advances in Neural Information Processing Systems*, *36*.

Huey, H., Jordan, M., Hart, Y., & Dillon, M. R. (2023). Mind-bending geometry: Children's and adults' intuitions about linearity on spheres. *Developmental psychology*, *59*(5), 886.

*Huey, H., *Oey, L. A., Lloyd, H., & Fan, J. E. (2023). How do communicative goals guide which data visualizations people think are effective? *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*(45).

Mukherjee, K., Lu, X., Huey, H., Vinker, Y., Aguina-Kang, R., Shamir, A., & Fan, J. E. (2023). Evaluating machine comprehension of sketch meaning at different levels of abstraction. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*(45).

Aboody, R., Huey, H., & Jara-Ettinger, J. (2022). Preschoolers decide who is knowledgeable, who to inform, and who to trust via a causal understanding of how knowledge relates to action. *Cognition*, *228*, 105212.

*Huey, H., *Long, B., Yang, J., George, K. R., & Fan, J. E. (2022). Developmental changes in the semantic part structure of drawn objects. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*(44).

Huey, H., Lu, X., Walker, C., & Fan, J. (2021). Explanatory drawings prioritize functional properties at the expense of visual fidelity.

Huey, H., Walker, C. M., & Fan, J. E. (2021). How do the semantic properties of visual explanations guide causal inference? *Proceedings of the Annual Conference of the Cognitive Science Society*, *43*(43).

Jara-Ettinger, J., Floyd, S., Huey, H., Tenenbaum, J. B., & Schulz, L. E. (2020). Social pragmatics: Preschoolers rely on commonsense psychology to resolve referential underspecification. *Child development*, *91*(4), 1135–1149.

Aboody, R., Huey, H., & Jara-Ettinger, J. (2018). Success does not imply knowledge: Preschoolers believe that accurate predictions reveal prior knowledge, but accurate observations do not. *Proceedings of the Annual Conference of the Cognitive Science Society*.

ABSTRACT OF THE DISSERTATION

Adaptive Visualization Strategies across Drawings, Diagrams, and Data Visualizations

by

Holly Huey

Doctor of Philosophy in Experimental Psychology

University of California San Diego, 2024

Professor Judith E. Fan, Co-Chair
Professor Caren M. Walker, Co-Chair

From Paleolithic etchings to computational graphics, visualization technologies that enable us to externalize our thoughts have been critical to the communication of ideas to others and across generations. Not only have renderings like simple line-drawings been prolific across cultures, but in recent years, generative text-to-visual systems have dramatically decreased production barriers and have increasingly populated our modern world with rich artifacts of visual communication. The study of *visual communication*—how people express their knowledge in visual form—presents abundant opportunities and challenges to explore core mechanisms of human communication systems, because

it relies on complex interactions between perception and pragmatics. *What guides what visualization strategies people use to convey their ideas to others through visual form?* My dissertation introduces new experimental methods to explore these strategies. Overall, I find that people's visual production behaviors shift what kind of information they prioritize depending on their task goals and find that their representational choices directly impact downstream interpretation by viewers. In Chapter 1, I investigate how communicative goals and immediate sensory inputs jointly determine the kind of visual information that people include in their drawings and find that people flexibly adapt their behavior to these different task goals by prioritizing different semantic information about visual object concepts. In Chapter 2, I explore how people adapt their visualization strategies when producing diagrams (called "visual explanations") of higher-level knowledge (e.g., object function vs. object appearance) and find that while these strategies facilitate inferences about physical mechanism, they do so at the expense visual fidelity and recognition of drawn objects. To evaluate whether such flexible visual communication behavior extends to more domain-specific knowledge, Chapter 3 explores how people evaluate what makes data visualization effective for different task goals and finds that people selectively prioritize different kinds of data visualization designs depending on what supports fast and accurate graph comprehension by viewers. Taken together, insights from these three lines of work build towards developing a more unified theory of visual communication and, ultimately, aim to better inform the development of human-centered visualization technologies guided by cognitive theories of visual communication.

Introduction

0.1 Visual communication

Our ability to understand pictures as semantically meaningful representations has long been contemplated by art historians (E. Gombrich, 1989) and philosophers (Greenberg, 2013). Records of pictorial representation predate evidence of written language by thousands of years (Aubert et al., 2014; Clottes, 2008; Hoffmann et al., 2018) and are ubiquitous across many cultures (E. H. Gombrich, 1960). However, despite the ancient and pervasive role that pictures have played throughout human history, relatively little is known about the cognitive mechanisms that underlie our ability to make and interpret pictures. In other words, how does the mind transform what are otherwise mere markings into interpretable representations?

The study of *visual communication*—how people express their knowledge in visual form—presents many opportunities and challenges to explore the core mechanisms of human communication, because it relies on interactions between perception and pragmatics. While a rich literature exploring "pictorial competence" (DeLoache & Burns, 1994) has offered insightful theoretical perspectives on how humans map the correspondence between visualizations and the referents that they represent, inconsistent behavioral evidence has left open key questions about what underlying processes are needed to understand visual representations[1]. On the one hand, because visualizations recapitulate many visual-spatial features of referents (e.g., in their shape, color, texture), a predominant perceptual theory is that our ability to interpret visualizations relies on the sophistication of our visual system's ability to recognize the resemblance between visual representations and the objects and scenes that they represent. Consequently, the success of visual production would rely on an ability to faithfully reconstruct those visual-spatial features in the form of,

[1] Pictorial competence encompasses a broad array of cognitive abilities, including perceiving, interpreting, producing, and using pictures. My dissertation focuses on production and interpretation and adopts a general definition of pictures: "A picture is constituted of marks on a delimited surface resulting from someone's attempt to communicate, preserve, or express an object." (DeLoache, et al., 1996, p. 3).

for example, drawings. On the other hand, growing research leveraging interactive dyadic contexts have instead shown evidence that our ability to understand visual representations is not only reliant on visual fidelity to object referents but instead can be learned from our interpretation of social cues through repeated interactions (J. E. Fan et al., 2018; Galantucci, 2005; Garrod et al., 2007). These studies have revealed important insight that while some visual representations may not be interpretable out of context, they are meaningful to people interacting with each other within social communication contexts. Insights from these studies suggest that our ability to produce recognizable visualizations is additionally reliant on our ability to use learned visual symbols (e.g., two dots and a c urve line to denote a smiley face) that others would understand. Towards reconciling these perspectives, my dissertation leverages modern psycholinguistics techniques, computational methods, and large-scale crowdsourcing to introduce new experimental methods aimed at: (1) exploring the visualization strategies that people use to convey different ideas to viewers; (2) evaluating how their visual production behavior may shift what information they prioritize in their visualizations, depending on their task goals; and (3) measuring how these representational choices may directly impact downstream interpretation by viewers. Insights from these works in my dissertation aim to help contribute towards a more unified theory of visual communication explaining how we encode and share our knowledge of the world with others through visual form.

0.2 Current theories of depiction based on *resemblance* vs. *convention*

Theories on the origins of pictorial competence have been dominated by two fiercely debated arguments. Several philosophers and notable vision scientists (e.g., James Gibson and colleagues) have argued that our perception of objects depicted in pictures is equivalent to our perception of real-world objects in the physical world (Gibson, 1978, 1971; J. M.

Kennedy, 1974). Intuitively, because pictures recapitulate visual-spatial features of their referents[2], they can be conceptualized as two-dimensional projections of three-dimensional objects (Cutting, 1986; Greenberg, 2021). Resemblance theories thus characterize pictures as mimetic surrogates that "*re*-present" (Danto, 1982) real-world objects through visual illusion. If the perception of objects depicted in pictures is reliant on the sophistication of the visual system to extract visual-spatial features of objects, regardless of whether they are real or represented, then picture perception may be an unlearned ability.

This claim resonates with behavioral evidence across developmental, cross-cultural, and comparative research on pictorial competence. Despite a lack of experience with pictorial representations, human infants (Hochberg & Brooks, 1962), some human adults living in cultures without pictorial art traditions or exposure to Western visual media (J. Kennedy, 1975), and higher non-human primates (Bovet & Vauclair, 2000; Tanaka, 2007) are able to recognize familiar objects in pictures. In particular, such recognition may critically depend on the tendency of pictures to capture the edge contours of objects through outlined form or differences in contrast. This may explain why line drawings tracing the outlines of photographed objects can drive object recognition so effectively (Biederman & Ju, 1988; Ishai et al., 2000). Additionally, recent computational modeling studies provide support that the perception of depicted objects relies on early visual processing. Deep convolutional neural networks of the visual cortex solely trained on natural photographs are able to recruit the same features in line drawings that support visual object recognition (J. E. Fan et al., 2018) and, conversely, convolutional neural networks trained on different types of pictures are able to recognize objects in natural photographs (Yin et al., 2016). Resemblance theories, thus, generally suggest that the perception of objects depicted in pictures relies on the same visual processing used for the perception of objects in the physical world.

By contrast, major critics of resemblance theories instead emphasize pictures as

[2] A "referent" is the real-world object depicted by the picture.

symbolic artifacts of human communication. Pictures under this account are considered to denote real-world objects through an abstract referential relation that necessarily depend on any visual resemblance. These conventionalist critics argue that pictorial competence is, instead, a *learned* ability that relies on the interpretation of social cues (Fay et al., 2010) and is influenced by cultural conventions (Goodman, 1976). Much in the same way that linguistic labels come to denote objects through association, observers must learn to associate pictures with real-world objects in order to learn that pictures denote those objects. For example, just as arbitrary linguistic labels like "bird" come to denote beaked and feathered egg-layers, so too might the "V" visual symbol eventually become associated with birds flying in sky scenes through cultural convention, despite bearing little resemblance to real-world birds (Goodman, 1976).

One strong interpretation is that observers *cannot* interpret pictures without first learning about the referential relationship between pictures and the real-world objects that they depict. This may explain why cross-cultural research has shown marked failures in recognition by other observers living in cultures without pictorial art traditions, ranging from a failure to recognize pictures as representational artifacts to a failure to recognize the objects depicted in those pictures at all (Deregowski, 1989). Simultaneously, this may also explain why observers of many Western cultures have a robust ability to understand highly schematized pictures, such as sketches produced by novices (J. E. Fan et al., 2018; Sangkloy et al., 2016) and abstract contemporary artwork, which both greatly diverge in fidelity from the visual-spatial features of their referents. Although prior accounts of pictorial competence have placed this strong interpretation in dichotomous contention against resemblance theories, recent compelling work has begun to disentangle how observers flexibly attend to *both* resemblance cues and intentional cues (Armitage, 2015; Armitage & Allen, 2015; Gelman & Ebeling, 1998). Consequently, a more moderate interpretation is that viewers' inferences about the mental state of the person who produced the picture guides their ability to extract the visual-spatial features of depicted objects necessary to recognize its referent. Therefore, instead of relying solely on the fidelity of the picture to help an observer map the visual-spatial features of the depicted object to the target referent, this interpretation contends that interpreting pictures rely on both faculties of social cognition and the sophistication of the visual system to extract those features of depicted objects.

This debate about pictorial competence has often been characterized as whether understanding pictures occurs with an "innocent eye" (E. Gombrich, 1969) or necessarily requires an "intelligent eye" (Gregory, 1970).

0.3 Evidence for depiction as a communicative act

How do we understand the communicative intent of other people? As an analogy, when we *talk* to others, the meaning of what we say is often revealed in the context of our conversations and shared knowledge. For example, if an individual asks their taller friend, "Can you please grab that book for me?" we can guess that the speaker is referring to a book tucked into an upper shelf and not a book on a lower shelf that the speaker can easily reach. As listeners, we are able to resolve such linguistic ambiguities by relying on our common sense psychology that enables us to form general expectations about other people's mental states and how they relate to their actions. While speakers must provide enough information so that listeners can recover the intended meaning of their utterances (Clark & Bangerter, 2004; Clark & Schaefer, 1987; Clark et al., 1983; Clark & Wilkes-Gibbs, 1986; Grice, 1975), listeners must infer what meaning is justified given what the speaker said. This interplay between speakers and listeners in verbal communication parallels a similar interplay between those producing visualizations and those viewing and interpreting them.

Research across a number of representational domains have shown that verbal and written language (Clark & Wilkes-Gibbs, 1986; Hawkins et al., 2017), novel sign languages Goldin-Meadow, 2005; Sandler et al., 2005, gesture (Goldin-Meadow et al., 1996), and drawing (Galantucci, 2005; Garrod et al., 2007) can emerge in short periods of time through repeated social interactions. This "creation through contact" account (Kegl, 1999) predicts that global communication systems first emerge through local interactions between—e.g., artists and observers in the same context—and morever, are not exclusive to linguistic domains. Although the identification of referents in pictures is more commonly used as a measurement of the success of a developed graphical sign to convey a discernible form, a few studies have begun to evaluate the contributions of a shared interaction history such as in repeated Pictionary-style communication games between paired participants.

Recent Pictionary-style studies have shown that adult sketchers, who are not allowed to speak or write letters or numbers in their drawings, produce increasingly sparse drawings across iterations of the same objects and scenes but converge on shared graphical signs by which to draw them (Garrod et al., 2007). Across iterations, participants' success with identifying the referents of their partner's drawings was shown to improve, although the drawings evolved from initially complex to simpler, more symbolic drawings. Moreover, participants tended to settle on drawing the same signs of referents (e.g., rabbit ears but not the entire rabbit) across iterations, suggesting that repeated interaction led participants to increased refinement of their pictures. However, when these simple signs were shown to naive observers, their comprehension was significantly lower compared to those who were shown the more complex signs generated in earlier iterations, demonstrating that social interaction was critical to communicative success in these contexts. Similar work has also demonstrated that paired participants in a series of visual communication games discovered increasingly sparse yet effective ways of depicting objects (Hawkins et al., 2019). Importantly, because the accuracy of observers' identification of target objects increased across iterations, these successful identifications were specific to the fact that target objects were repeated referenced by sketchers in the task, rather than accurate fidelity to the referents. Furthermore, while the visual features of drawings within pairs became increasingly consistent, other pairs of participants markedly diverged in the graphical signs they developed to convey the same objects, demonstrating the emergence of these sparse pictures were determined from local interactions between sketchers and observers of shared knowledge.

These studies offer critical initial insights about how the variability and expressiveness of drawing behavior, as an instance of visual communication, is nonetheless systematic depending on the context in which the drawings are being produced. However, these works have often predominately leveraged drawing to investigate how people communication object identity by appearance and have often heavily relied on manually annotating drawings, which has inherently limited the sample sizes necessary to explore large-scale patterns in visual production behavior. Other studies have leveraged modern deep learning models to characterize feature distinctiveness of drawings at scale (J. Fan et al., 2018; Long et al., 2024; Yamins et al., 2014), but have been limited to investigation of object appearance because such models cannot explicitly encode semantic part-level information about drawn objects. My dissertation builds on and aims to broaden the scope of these investigations on visual communication across a wide array of communicative contexts and diverse visualizations: drawings, diagrams, and data visualizations.

0.4 Current work

Throughout my dissertation, I introduce new experimental methods to evaluate the strategies that people leverage to convey their knowledge through visual form to others in different communicative contexts. Specifically, my research measures the degree which different task goals systematically shift visual production behavior depending on task goals and how differences in representational choices impact downstream interpretation by viewers. One null hypothesis is that because visual communication spans a wide and flexible range of visual expressivity that people do not demonstrate systematic visual production behavior across task goals. By contrast, people's visual production behavior may systematically quite narrow across task goals if their main goal is to visually reconstruct object referent to their best of their ability. This second hypothesis resonates with resemblance theories of depiction.

However, across the three chapters of my dissertation, I demonstrate that people's visual production behavior is neither solely driven by individual differences in style nor goals to visually reconstruct the appearance of the object and scenes in their environments. Rather, my work provides evidence that visual communication is a cognitive act that balances portraying what is visually salient and semantically relevant to viewers, guided by pragmatic inferences about what a viewer may need to gain from the visualization depending on their task goals. Two main predictions arise from this idea: one possibility, which I call the *cumulative* hypothesis, is that people increasingly add visual information to their representations as they scale different levels of abstraction required by their tasks. Another possibility, which I call the *dissociable* hypothesis, is that people may selectively represent different aspects of object concepts such that different types of visual information may trade off with one another. I explore each of these hypotheses through drawings of visual object concepts, diagrams of higher-level knowledge (called "visual explanations"), and data visualizations in the following chapters.

In Chapter 1, I investigate how communicative goals and immediate sensory inputs jointly determine the kind of visual information that people include in their drawings and find that people flexibly adapt their behavior to these different task goals by prioritizing different semantic information about visual object concepts. In Chapter 2, I explore how people adapt their visualization strategies when producing visual explanations of higher-level knowledge (e.g., object function vs. object appearance) and find that while these strategies facilitate inferences about physical mechanism, they do so at the expense visual fidelity and recognition of drawn objects. To evaluate whether such flexible visual communication behavior might extend to more domain-specific experience, Chapter 3 explores how people evaluate what makes data visualization effective for different task goals and find that people selectively prioritize different kinds of data visualization designs of the same data depending what would support graph comprehension by viewers.

Overall, I find that people systematically shift how they choose to visually represent their knowledge depending on their task goals and moreover, selectively prioritize different kinds of visual information. On the whole, these findings support a *dissociable* hypothesis and show evidence for the remarkable and creative ways in which people adaptively shift their visual production behaviors. These findings also show that these representational choices to prioritize different visual and semantic information lead to measurable differences in how well viewers can interpret these visualizations depending on their own task goals. Taken together, the results of my dissertation research aim at contribute toward developing more unified theories of visual communication. In the long run, these insights aim to support the creation and improvement of human-centered visualization technologies informed by cognitive theories of visual communication.

References

Armitage, E. (2015). *A theory of pictures: Investigating the mediating role of picture modality in children's understanding of the picture-creator and picture-referent relationships* (Doctoral dissertation). Lancaster University.

Armitage, E., & Allen, M. L. (2015). Children's picture interpretation: Appearance or intention? *Developmental psychology*, *51*(9), 1201.

Aubert, M., Brumm, A., Ramli, M., Sutikna, T., Saptomo, E. W., Hakim, B., Morwood, M. J., van den Bergh, G. D., Kinsley, L., & Dosseto, A. (2014). Pleistocene cave art from sulawesi, indonesia. *Nature*, *514*(7521), 223–227.

Biederman, I., & Ju, G. (1988). Surface versus edge-based determinants of visual recognition. *Cognitive psychology*, *20*(1), 38–64.

Bovet, D., & Vauclair, J. (2000). Picture recognition in animals and humans. *Behavioural brain research*, *109*(2), 143–165.

Clark, H. H., & Bangerter, A. (2004). Changing ideas about reference. In *Experimental pragmatics* (pp. 25–49). Springer.

Clark, H. H., & Schaefer, E. F. (1987). Collaborating on contributions to conversations. *Language and cognitive processes*, *2*(1), 19–41.

Clark, H. H., Schreuder, R., & Buttrick, S. (1983). Common ground at the understanding of demonstrative reference. *Journal of verbal learning and verbal behavior*, *22*(2), 245–258.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39.

Clottes, J. (2008). *Cave art*. Phaidon London.

Cutting, J. E. (1986). *Perception with an eye for motion* (Vol. 1). Mit Press Cambridge, MA.

Danto, A. C. (1982). Depiction and description. *Philosophy and phenomenological research*, *43*(1), 1–19.

DeLoache, J. S., & Burns, N. M. (1994). Early understanding of the representational function of pictures. *Cognition*, *52*(2), 83–110.

Deregowski, J. B. (1989). Real space and represented space: Cross-cultural perspectives. *Behavioral and Brain Sciences*, *12*(1), 51–119.

Fan, J., Yamins, D., & Turk-Browne, N. (2018). Common object representations for visual production and recognition. *Cognitive Science*, *42*, 2670–2698.

Fan, J. E., Yamins, D. L., & Turk-Browne, N. B. (2018). Common object representations for visual production and recognition. *Cognitive science*, *42*(8), 2670–2698.

Fay, N., Garrod, S., Roberts, L., & Swoboda, N. (2010). The interactive evolution of human communication systems. *Cognitive science*, *34*(3), 351–386.

Galantucci, B. (2005). An experimental study of the emergence of human communication systems. *Cognitive science*, *29*(5), 737–767.

Garrod, S., Fay, N., Lee, J., Oberlander, J., & MacLeod, T. (2007). Foundations of representation: Where might graphical symbol systems come from? *Cognitive science*, *31*(6), 961–987.

Gelman, S. A., & Ebeling, K. S. (1998). Shape and representational status in children's early naming. *Cognition*, *66*(2), B35–B47.

Gibson, J. J. (1978). The ecological approach to the visual perception of pictures. *Leonardo*, *11*(3), 227–235.

Gibson, J. J. (1971). The information available in pictures. *Leonardo*, *4*(1), 27–35.

Goldin-Meadow, S. (2005). *The resilience of language: What gesture creation in deaf children can tell us about how all children learn language*. Psychology Press.

Goldin-Meadow, S., McNeill, D., & Singleton, J. (1996). Silence is liberating: Removing the handcuffs on grammatical expression in the manual modality. *Psychological Review*, *103*(1), 34.

Gombrich, E. (1989). *The story of art*. Phaidon Press, Ltd.

Gombrich, E. (1969). *Art and illusion: A study in the psychology of pictorial representation*. Princeton, NJ: Bollingen Series/Princeton University Press.

Gombrich, E. H. (1960). *A study in the psychology of pictorial representation*. Pantheon Books.

Goodman, N. (1976). *Languages of art: An approach to a theory of symbols*. Hackett publishing.

Greenberg, G. (2013). Beyond resemblance. *Philosophical review*, *122*(2), 215–287.

Greenberg, G. (2021). Semantics of pictorial space. *Review of Philosophy and Psychology*, 1–41.

Gregory, R. L. (1970). The intelligent eye.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and Semantics Vol. 3: Speech Acts* (pp. 64–75). Academic Press.

Hawkins, R. X., Frank, M., & Goodman, N. D. (2017). Convention-formation in iterated reference games. *CogSci*.

Hawkins, R. X., Sano, M., Goodman, N. D., & Fan, J. W. (2019). Disentangling contributions of visual information and interaction history in the formation of graphical conventions. *CogSci*, 415–421.

Hochberg, J., & Brooks, V. (1962). Pictorial recognition as an unlearned ability: A study of one child's performance. *The American Journal of Psychology*, *75*(4), 624–628.

Hoffmann, D. L., Standish, C. D., García-Diez, M., Pettitt, P. B., Milton, J. A., Zilhão, J., Alcolea-González, J. J., Cantalejo-Duarte, P., Collado, H., de Balbín, R., Lorblanchet, M., Ramos-Muñoz, J., Weniger, G.-C., & Pike, A. W. G. (2018). U-th dating of carbonate crusts reveals neandertal origin of iberian cave art. *Science*, *359*(6378), 912–915. https://doi.org/10.1126/science.aap7778

Ishai, A., Ungerleider, L. G., Martin, A., & Haxby, J. V. (2000). The representation of objects in the human occipital and temporal cortex. *Journal of Cognitive Neuroscience*, *12*(Supplement 2), 35–51.

Kegl, J. (1999). Creation through contact: Sign language emergence and sign language change in nicaragua. *Language creation and language change: Creolization, diachrony, and development*.

Kennedy, J. (1975). Drawings were discovered, not invented. *New Scientist*, *67*(523), 764.

Kennedy, J. M. (1974). *A psychology of picture perception*. Jossey-Bass Publishers.

Long, B., Fan, J. E., Huey, H., Chai, Z., & Frank, M. C. (2024). Parallel developmental changes in children's production and recognition of line drawings of visual concepts. *Nature Communications*, *15*(1), 1191.

Sandler, W., Meir, I., Padden, C., & Aronoff, M. (2005). The emergence of grammar: Systematic structure in a new language. *Proceedings of the National Academy of Sciences*, *102*(7), 2661–2665.

Sangkloy, P., Burnell, N., Ham, C., & Hays, J. (2016). The sketchy database: Learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)*, *35*(4), 1–12.

Tanaka, M. (2007). Recognition of pictorial representations by chimpanzees (pan troglodytes). *Animal cognition*, *10*(2), 169–179.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, *111*(23), 8619–8624.

Yin, R., Monson, E., Honig, E., Daubechies, I., & Maggioni, M. (2016). Object recognition in art drawings: Transfer of a neural network. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2299–2303. https://doi.org/10.1109/ICASSP.2016.7472087

Chapter 1

Visual communication of object concepts at different levels of abstraction

Abstract

Visual communication—how people flexibly express their knowledge in visual form—has been critical to human creativity and collaboration. Here we investigate how people adaptively prioritize semantic features of visual object concepts in their drawings depending on their task context: (1) when given goals to convey different levels of object information, and (2) when target objects are cued through visual experience or semantic memory. In Experiment 1, we elicited >12K drawings of 32 object categories by presenting participants with photos or text labels and asking them to draw a general object category or a specific exemplar of it. In Experiment 2, we acquired two recognizability scores for each drawing of our dataset. In a categorization task, naive viewers guessed the object category of a drawing, providing a measure of the category-diagnostic information contained within it. In an identification task, another group of naive viewers matched each photo-cued drawing to a corresponding photo, providing a measure of its exemplar-diagnostic information. Lastly in Experiment 3, we collected dense annotations of how every stroke of the drawings corresponded to parts of the target object categories and used them to evaluate an array of semantic features in the drawings, as well as to measure their impact on downstream recognizability by viewers. We found that category-goal drawings included less visual detail but were more categorizable to viewers. By contrast, exemplar-goal drawings included more visual detail and object parts, but were least categorizable to viewers. However, photo-cued exemplar-goal drawings were more often correctly matched to photos. Taken together, our findings suggest that people flexibly adapt their visual production behaviors to different task contexts by prioritizing different semantic information about visual object concepts.

Keywords: sketch production; sketch recognition; visual object concepts; semantic cognition

1.1 Introduction

Throughout human history, visualization technologies (e.g., maps, diagrams, data visualizations) have been an essential modality through which people share and acquire knowledge about the objects in our environments. In particular, line drawings remain among the most enduring and expressive techniques for capturing key visual properties of objects diagnostic of their category membership (J. E. Fan et al., 2018), including granular detail such as what parts an object is composed of and how they are arranged to make the object what it is (Lu et al., 2023). This expressiveness of drawing has enabled researchers across cognitive and developmental psychology, linguistics, and computer vision to use drawings as a rich case study for investigating adaptive human communication—for example, when drawing across repeated dyadic interactions (R. D. Hawkins et al., 2023), in tasks with limited time or available pen strokes (Mukherjee et al., 2024), conveying different semantic information (Holt et al., 2024; Huey et al., 2023), or distinguishing between objects of the same or different categories (Mukherjee et al., 2019). These studies have frequently found that successful visual communication primarily depends on how effectively people can adapt to task conditions and distill the relevant perceptual information of objects in their drawings for others. While the last several years have seen remarkable progress in our understanding of the flexibility of visual production behavior, current theories of depiction have nonetheless fallen short of systematically explaining how people arbitrate between conveying high-level knowledge about general object categories (e.g., “bird”, “flower”, “bread”) and more specific knowledge about object exemplars (e.g., “sparrow”, “daisy”, “focaccia”). This leaves open critical questions about how the salience of certain object properties may shift under different task contexts, as well as how people map the correspondence of the semantic information represented in their drawings to the features of real-world objects that are diagnostic of their category membership.

This challenge is driven by methodological differences across current drawing paradigms that often use visual or linguistic cues independently to prompt drawings of object categories, but is also rooted in core questions about how people draw upon their immediate visual experience (e.g., via photos that show specific object exemplars) or semantic memory (e.g., via text labels of object categories). These questions build on natural language theories that highlight that our semantic representations of objects can span multiple levels of abstraction. For example, while people tend to categorize objects at a basic level (G. L. Murphy & Smith, 1982; E. H. Rosch, 1973), people can also produce different labels for the same object (e.g., subordinate level: "Garfield", basic level: "cat", superordinate level: "animal"), depending on what is informative for a particular context (Degen et al., 2020). Moreover, this ability to discriminate objects at finer levels of granularity can improve with the acquisition of expertise (Tanaka, 2001; Tanaka & Taylor, 1991), suggesting that goal-directed and sustained attention to diagnostic features of objects can lead to changes in the accessibility of concepts at different levels of abstraction (Nosofsky, 1986). However, while prior research demonstrates that such knowledge can be evoked through multiple modalities, such as through visual and linguistic cues (Potter, 1976), the majority of studies aimed at characterizing object category representations through part evaluations have typically leveraged linguistic cues (e.g., text labels) to access semantic memory (Garrard et al., 2001; McRae et al., 2005; G. Murphy, 2004; E. Rosch et al., 1976; Tversky & Hemenway, 1984). By contrast, visual cues (e.g., photos) have predominately been used to approximate our visual experience and are often used to prompt broad categorizations of objects (Biederman, 1987; Edelman et al., 1999; G. L. Murphy & Smith, 1982; E. H. Rosch, 1973). Seminal work on mental imagery has shown that people are faster to verify larger or smaller parts of animal categories (e.g., "head" vs. "claws" of a cat) depending on whether they are presented with a photo cue or not, even if smaller parts are often more strongly associated with the animal category (e.g., "whiskers") (Kosslyn, 1976a, 1976b), suggesting that access to visual experience can shift the salience of object parts to viewers. However, given the methodological differences in cue modality across investigations on categorical knowledge, it remains unclear what semantic features of object categories may be more or less salient to people when producing drawings meant to convey categorical-level and exemplar-level information to others.

Because objects can be parsed into their constituent parts based on their perceptual and functional salience (e.g., a “cat” has a head, whiskers, tail, paws. . .), foundational research in semantic cognition has often focused on evaluating *partonomic knowledge* as a way to probe how people conceptually represent the semantic properties and structure of object categories (Tversky & Hemenway, 1984). As children acquire more knowledge about the objects in their environments, their drawings become more visually distinctive (Long et al., 2024) and increasingly inclusive of diagnostic object properties across childhood (Barrett & Light, 1976; Bremner & Moore, 1984; Sitton & Light, 1992). Conversely, adults’ drawings that fail to include features that are diagnostic of target object categories have been associated with semantic dementia (Bozeat et al., 2003; Rogers & Patterson, 2007). These studies corroborate robust linguistic research showing that people understand that objects possess individuating features that indicate their membership within broader categories (McClelland & Rumelhart, 1985; Palmeri & Gauthier, 2004). However, while these developmental and clinical insights emphasize that drawings encode more than immediate visual experience but also conceptual knowledge, they traditionally rely on manual annotations of drawings of singular object categories (e.g., mugs, people, ducks). These approaches not only limit samples sizes which cannot fully explore adaptive visual productive behavior across different contexts, but also limit part-level analyses to object parts predetermined by researchers. Additionally, while recent visualization research efforts have developed innovative techniques for computationally analyzing drawing features across large-scale datasets spanning numerous object categories (Mukherjee et al., 2024), these investigations have focused on evaluating drawing features through high-level recognizability of object categories or low-level effort metrics, like analyzing stroke count. Simultaneously, modern deep learning models have made critical progress towards characterizing feature distinctiveness of object categories (Dosovitskiy et al., 2020; He et al., 2016; Radford et al., 2021; Simonyan & Zisserman, 2014; Yamins et al., 2014), including sketch categorization (Ballester & Araujo, 2016; Eitz, Hays, et al., 2012; Yu et al., 2017), segmentation (Li et al., 2018; Yang et al., 2021), and shape retrieval (Bhunia et al., 2020; Eitz, Richter, et al., 2012; Sangkloy et al., 2016; Song et al., 2017; Yu et al., 2016), but cannot yet explicitly encode part-level recognition of object categories.

Towards tackling these challenges, our work generates a large-scale dataset of semantically annotated drawings spanning 32 diverse object categories in order to investigate how people prioritize different parts of visual object concepts depending on (1) their communicative goals to convey general category-level information or specific exemplar-level information, and (2) on how target categories are accessed via visual experience or semantic memory. To do this, we collected >12K densely annotated drawings by varying whether they were produced for a specific object exemplar or broad object category and whether they were evoked by photo or text label cue. Next, we developed a series of distinct measurements to evaluate three distinct sketch features: (1) *part complexity* measuring the inclusion of unique object parts, (2) *part emphasis* measuring the number of strokes allocated to each part, and (3) *part spatial arrangement* measuring the spatial layout and compositionality of those parts (Fig. 1.5). By measuring these drawing features at the part-level, we could then estimate their impact on their downstream recognizability by viewers. One possibility is that visual features of a specific exemplar are integrated with the more general visual features about the broad object category it belongs to. Under this account, when people have the goal of producing a drawing that has a more specific meaning (e.g., a particular cat like "Garfield"), they may include features that are not only diagnostic at the exemplar level, but also at the category level. Alternatively, it may be the case that the visual features diagnostic of a specific exemplar are dissociable from visual features of the broad object cateogry it belongs to. This alternative account would predict that while drawings intended to communicate more specific meanings may contain information that is diagnostic of object *identity*, they may not necessarily communicate *category* information as effectively. Finally, the degree to which people include information at both levels of abstraction may depend on whether they have immediate visual access to an object's appearance (e.g., when prompted with a photo cue), or whether they are relying exclusively on semantic memory (e.g., when prompted with a text label).

By evaluating the semantic part features that people adaptively prioritize in their visual production, our work aims to explore how communicative goals and immediate sensory inputs jointly determine what kind of visual information people externalize in their drawings of visual object concepts. Our work builds on recent efforts to semantically analyze drawing datasets that span multiple object categories at the part-level (Long et al., 2024) as well as growing efforts to develop large-scale drawing datasets aimed at exploring visual production behavior (Dey et al., 2019; Eitz, Hays, et al., 2012; J. E. Fan et al., 2023; Jongejan et al., 2017; Li et al., 2018; Mukherjee et al., 2023; Sangkloy et al., 2016), but is the first dataset, to our knowledge, that systematically manipulates task context. By publicly releasing our dataset and proposed evaluation protocol for investigating adaptive visual production behavior, we hope that these data will provide new opportunities for future research seeking to understand how we flexibly encode and externalize our knowledge of our visual world when communicating with others.

1.2 Experiment 1: Sketch Production

Our first goal was to generate a large-scale sketch dataset in order to investigate how people adapt their visual production behavior across task contexts. To accomplish this, we asked people to produce sketches that represent the general idea of a target concept (i.e., categorical information) or represent a specific instance of a target concept (i.e., exemplar information), as well as manipulated whether the sketch was cued by a photo vs. text label of a concept. By independently manipulating the goal and cue type of drawing, we could then estimate the impact of these task constraints on sketch production behavior.

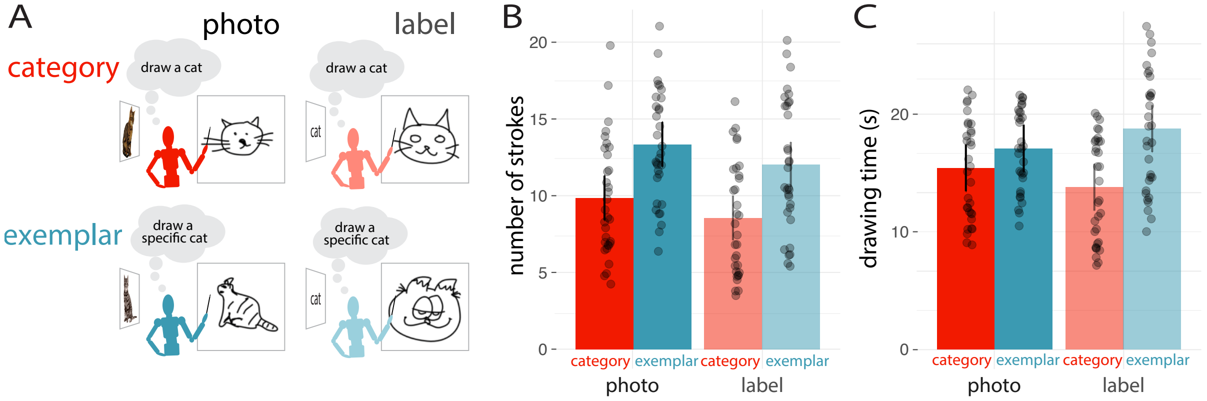


Figure 1.1. (A) Sketch production task. Each participant produced 32 drawings of different object, with one of two goals (exemplar vs. category) and cue-type conditions (photo vs. label). (B) Number of strokes used to produce each sketch. (C) Amount of time spent producing each sketch. Error bars reflect 95% CIs.

1.3 Method

1.3.1 Sketch production experiment

1.3.2 Participants

We recruited 384 participants (128 female, 25.9 years) to participate in our study via Prolific. Each participant received $6.00 for their participation (approx. $12/hr). We did not exclude data from any participant, as none met our pre-registered exclusion criteria. [1]

1.3.3 Stimuli

We designed our stimulus set to build upon existing benchmark datasets in computer vision containing digital drawings of real-world visual objects (Eitz, Hays, et al., 2012; Sangkloy et al., 2016). Out of the 125 object categories in the *Sketchy* database (Sangkloy

[1] The pre-registered exclusion criteria for this and subsequent experiments can be found here: https://github.com/cogtoolslab/photodraw32/blob/master/plan/photodraw_preregistrations.txt

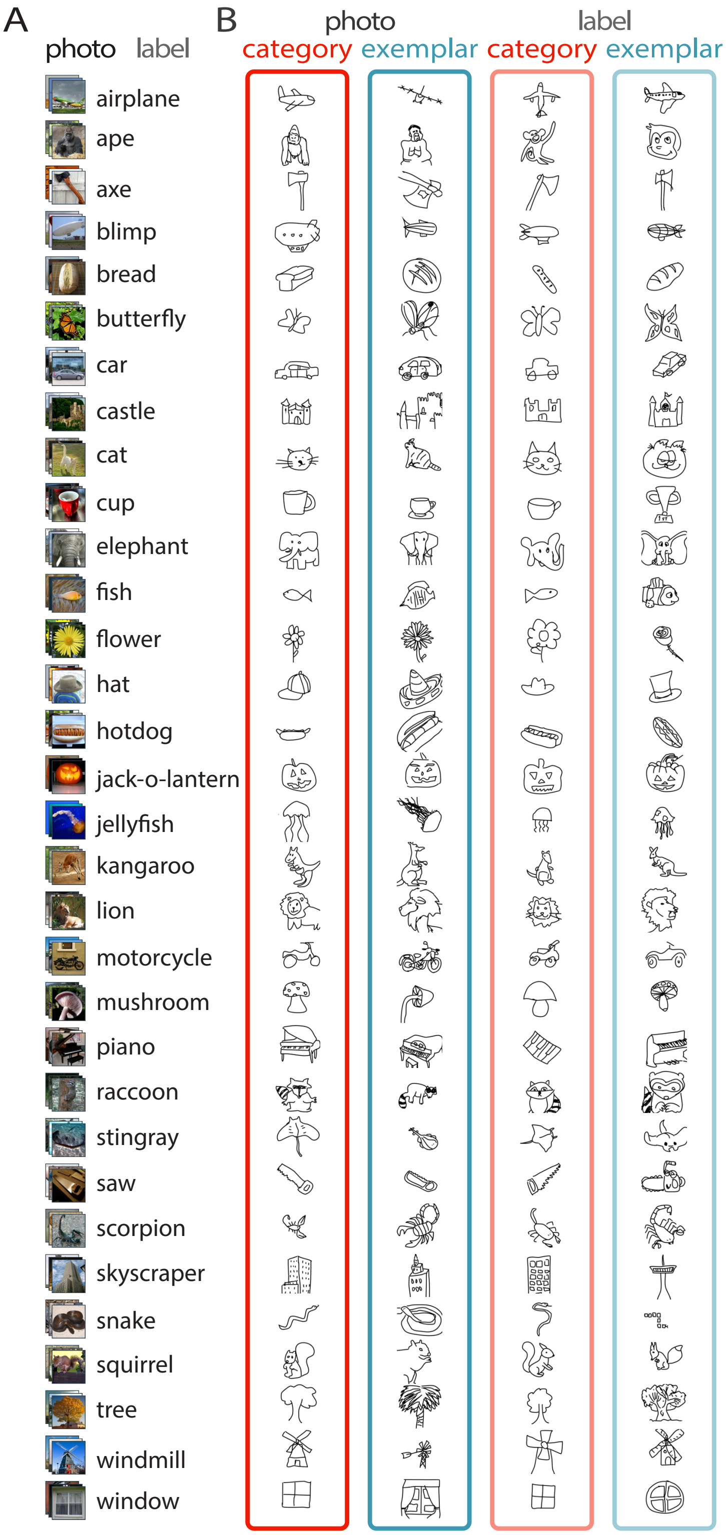


Figure 1.2. Sketch examples across goal and cue conditions

et al., 2016), we selected 32 categories spanning a wide range of familiar concepts and approximately balanced with respect to animacy (living/nonliving), size (large/small), familiarity (high/low), and naturalness (natural/artificial). These categories were: *airplane, ape, axe, blimp, bread, butterfly, car (sedan), castle, cat, cup, elephant, fish, flower, hat, hotdog, jack-o-lantern, jellyfish, kangaroo, lion, motorcycle, mushroom, piano, raccoon, ray, saw, scorpion, skyscraper, snake, squirrel, tree, windmill, and window.* Then, out of the 100 photographs from each category in the *Sketchy* database, we selected 32 exemplar images that varied with respect to both category-orthogonal properties (e.g., pose, viewpoint) as well as category-relevant properties (e.g., typicality).

1.3.4 Task Procedure

To generate a large-scale sketch dataset, we designed a web-based drawing platform in which participants were prompted to produce different object concepts and could use this cursor to draw on an virtual canvas. We manipulated cue type and communicative goal using a 2x2 between-participants design, such that each participant was pseudo-randomly assigned to a cue-type (i.e., photo vs. text label) and abstraction-level (i.e., exemplar vs. category) condition (Fig. 1.1A; N=96 participants per condition). Participants in the photo-exemplar group were instructed to: "make a drawing that would help someone else looking only at your drawing guess which image you were prompted with out of a lineup containing other similar images." By contrast, participants in the photo-category group were instructed to: "make a drawing that is recognizable, but not one that could be matched to the image I was shown." Participants in the label-category group were instructed to: "make a drawing that would help someone else looking only at your drawing guess which word you were prompted with." Finally, participants in the label-exemplar group were instructed to visualize and "draw a *specific* object, rather than a general object category."

Each participant produced 32 drawings, one per category. To equate the total amount of preparation time participants in all four groups had before beginning their drawing, the cue was always presented for 8 seconds and then removed before participants could begin their drawing. The sequence in which categories appeared across trials was randomized across participants, but the number of times a given photo was presented was balanced, such that each photo served as the cue three times across all photo-cue experimental sessions. Our resulting dataset contained 12,288 drawings with 382 drawings per category (Fig. 1.2). At the end of each session, participants were asked to report their own level of drawing skill ("How skilled do you consider yourself to be at drawing?") using a 7-point Likert scale.

1.4 Results

Based on prior work (J. E. Fan et al., 2020), we hypothesized that drawings emphasizing distinctions between exemplars would be more detailed than those that only needed to be recognizable at the category level in order to produce sufficiently informative drawings at each level of abstraction. We further hypothesized that, insofar as the photo cues served to activate participant's memories of what objects visually looked like, that photo-cued drawings would also be more detailed than label-cued ones.

1.4.1 How much detail do people represent in their drawings depending on their task context?

To evaluate these hypotheses, we analyzed the number of strokes and amount of time participants used to produce each sketch. We first fit a linear mixed-effects model to predict the number of strokes from communicative goal and cue type, with random intercepts for each participant and category, and found that this model outperformed nested variants of this model containing communicative goal ($\chi^2 = 9.8$, $p = 1.7\text{e}{-3}$) or cue type ($\chi^2 = 79.0$, $p < 1\text{e}{-16}$) alone. However, adding an interaction term between cue type and communicative goal did not improve model fit ($\chi^2 = 0.11$, $p = 0.73$), suggesting that each task factor independently impacted the number of strokes participants included in their drawings. Specifically, participants who had the goal of communicating exemplar-level information used more strokes relative to those aiming to convey a category (exemplar: 12.7 strokes, 95% CI: [11.3, 14.1]; category: 9.20 strokes, 95% CI: [7.79, 10.6]; $b = 3.49$, $t_{376} = 9.35$, $p <$ 2e$-$16; Fig. 1.1B). Moreover, participants who were cued with a photo used more strokes than those cued with a category label (photo: 11.6 strokes, 95% CI: [10.2, 13.0]; label: 10.3 strokes, 95% CI: [8.86, 11.7]; $b = 1.30$, $t_{376} = 3.49$, $p =$ 5.47e$-$3).

Next, we applied the same procedure to model the amount of time participants spent producing each drawing. Here we found that a mixed-effects model containing communicative goal, cue type, and their interaction as predictors outperformed a nested variant lacking the interaction term ($\chi^2 = 6.22$, $p = 0.0126$). Further inspection of the coefficients of each term revealed that having the exemplar goal (exemplar: 17.9 seconds, 95% CI: [16.2, 19.7]; category: 14.6 seconds, 95% CI: [12.8, 16.4]; $b = 1.66$, $t_{378} = 1.76$, $p = 0.0788$) and being cued with a photo led to modest increases in the amount of time participants spent producing their drawing (photo: 16.3 seconds, 95% CI: [14.5, 18.0]; label: 16.3 seconds, 95% CI: [14.5, 18.0]; $b = 1.65$, $t_{378} = 1.75$, $p = 0.0808$), with the effect of communicative goal being larger in the label-cue context (exemplar: 18.8 seconds, 95% CI: [16.8, 20.8]; category: 13.8 seconds, 95% CI: [11.8, 15.8]; $b = 3.33$, $t_{378} = 2.49$, $p = 0.0130$; Fig. 1.1C).

Together these findings provide converging evidence that both the target communicative goal and the availability of a visual cue impact the level of detail in participants' drawings of these visual object concepts, replicating and extending prior work that had only manipulated communicative goals (J. E. Fan et al., 2020).

1.5 Experiment 2: Sketch Recognition

Our findings so far suggest that what information people have visual access to (photo vs. label) and what information they aim to convey (category vs. exemplar) impact how much time and effort they spend producing their drawing. Our next goal was to evaluated whether these differences in drawing task context lead to meaningful difference in recognizability of target concepts, beyond mere stylistic differences in the drawings. In Experiment 2, we accomplish this by conducting two recognition tasks, asking a new group of naive participants to identify which target category a drawing represented and another group of participants to identify which specific photo exemplar a drawing represented. We also conducted a photo evaluation task to develop measures of how well participants thought the photo stimuli represented the target categories, in order to evaluate how difference in visual information in photos may lead to more or less recognizability of categorical or exemplar information in the corresponding sketches.

We hypothesized that, insofar as people intending to produce a drawing of an exemplar also include category-diagnostic information, drawings produced under either communicative goal would evoke the target category to a similar degree. For instance, suppose that when someone intends to produce a drawing of a specific cat, they aim to approximate its visual appearance as closely as possible. So long as the drawing looks like *that* cat, it may contain the relevant information to drive both its categorization as a cat and identification of the specific cat. Alternatively, different information may be prioritized when communicating category vs. exemplar information. For example, there may be distinctive visual features that can be used to help identify the specific cat (e.g., unusually large eyes) and distinguish it from other cats, but these features are not necessarily strongly diagnostic of category membership. Furthermore, supposing that these distinctive features are more prominent for less prototypical exemplars, we also examined the relationship between the typicality of an photo cue and how easily the resulting drawings could be recognized at the category and exemplar levels.

1.6 Method

1.6.1 Sketch categorization task

1.6.2 Participants

347 participants (104 female, mean age = 24.8 years) were recruited via Prolific and received $3.00 for their participation (approx. $12/hr). Of these participants, 15 dropped out of the study early and one participant failed to meet our predefined exclusion criteria. The data from these 16 sessions were excluded from further analysis, yielding 331 complete sessions.

1.6.3 Task procedure

On each trial, participants were presented with a drawing from Experiment 1 and asked to select the category label that best matched it from among the full set of 32 category labels used in Experiment 1 (*"Which category does this drawing belong to?"*; Fig. 1.3A). Each categorization participant completed 128 trials, excluding 4 catch trials. Overall, this procedure yielded 42,368 judgments, with each drawing having received at least 3 categorization responses.

1.6.4 Sketch exemplar identification task

1.6.5 Participants

160 participants (36 female, mean age = 24.3 years) were recruited via Prolific to complete a recognition task in which they were shown a photo-cued drawing and were asked to match the drawing to one of 8 possible images. Each participant received $5.00 for their participation in the 25-minute study (approx. $12/hr). Of these participants, three dropped out of the study early and one failed to meet our preregistered exclusion criteria. The data from these four sessions were excluded from further analyses.

1.6.6 Task procedure

Each session consisted of 128 trials. On each trial, participants were presented with one of the drawings from Study 1 and eight color photos and were instructed to select the photo they believed the drawing was intended to depict (Fig. 1.3D). One of these photos was always the actual image cue used to elicit the drawing in Experiment 1, and the other seven distractor images were sampled from among the remaining 31 photos belonging to the same category. To place sufficient demands on exemplar-level discrimination, we identified the seven images that were most visually similar to the actual cue using an automated procedure. Specifically, we estimated the perceptual similarity between every pair of photos in our stimulus set by calculating the correlation between feature-vector representations of each image computed by a deep convolutional neural network that had been trained on a large and independent set of images (Deng et al., 2009; Simonyan & Zisserman, 2014). We then found the seven images whose feature vectors had the highest correlation (i.e., nearest neighbors) with that of the actual cue. We additionally included four catch trials containing especially identifiable drawings that we expected any sufficiently engaged participant to be able to successfully identify. These catch trials were interleaved among the other drawings and identical for all participants.

1.6.7 Photo typicality evaluation task

1.6.8 Participants

88 participants (42 male, mean age = 29.2 years) were recruited via Prolific. Each participant received \$3.00 for their participation in the 15-minute study (approx. \$12/hr). Data from 8 participants who did not meet the exclusion criteria we defined after data collection (but prior to formal analysis) were excluded from further analyses.[2] Overall, we

[2] Data from an entire session were excluded if: (1) four or more catch trials out of eight were failed, (2) there were two or more response "streaks" wherein the same rating was given eight times in succession, and (3) the pattern of ratings across trials had an unusually low correlation with ratings provided by other participants.

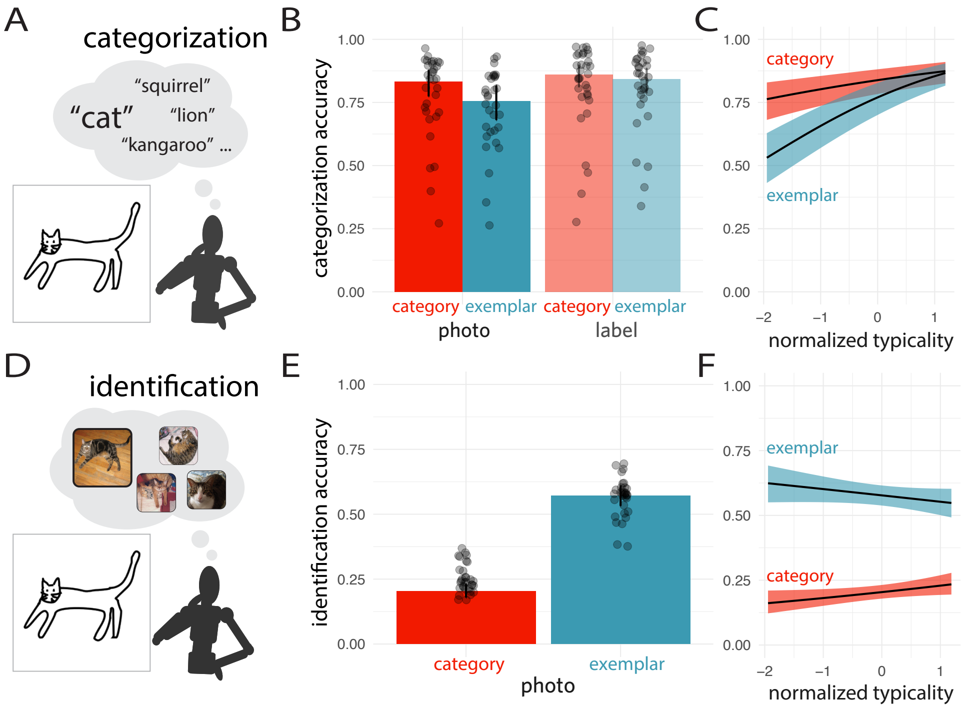


Figure 1.3. (A) Sketch categorization task. Participants selected the category label that best matched the drawing from the full set of 32 category labels. (B) Proportion of drawings correctly categorized across conditions. Error bars reflect 95% CIs. (C) Categorization accuracy for photo-cued drawings as a function of the typicality of the photo cue. Error ribbons reflect 95% CIs. (D) Sketch exemplar identification task. Participants selected the photo that best matched the drawing from a set of 8 photos belonging to the target category. Only photo-cued drawings were included in this experiment. (E) Proportion of photo-cued drawings correctly matched to their corresponding photo cue. (F) Exemplar identification accuracy for photo-cued drawings as a function of the typicality of the photo cue.

discovered that small adjustments to these exclusion criteria did not have a major impact on our key analyses.

1.6.9 Task procedure

Each participant was presented with the prompt (*”How well does this picture fit your idea or image of the category?”*), a series of 128 images, and was asked to provide typicality judgments on a 5-point Likert scale: “Not at all”, “Somewhat”, “Moderately”, “Very”, and “Extremely.” In each session, there were 4 images from each of the 32 categories. This study yielded 10,240 ratings, with each photo receiving 10 ratings each.

1.7 Results

1.7.1 How does communicative goal and cue type impact category-level recognition?

We first sought to measure the impact of communicative goal and cue type on the amount of category evidence in a drawing, as measured by how well naive participants could determine the category a drawing was intended to convey. We used a mixed-effects logistic regression model to predict the outcome of each categorization judgment using communicative goal, cue type, as well as their interaction, with random intercepts for drawing participant, categorization participant, and target category. This model outperformed nested variants that did not include the interaction term ($\chi^2 = 7.13$, $p = 0.008$) or excluded either communicative goal ($\chi^2 = 11.7$, $p = 6.3\text{e}{-4}$) or cue type ($\chi^2 = 28.7$, $p = 8.5\text{e}{-8}$), suggesting that both task factors influenced the categorizability of a drawing. To control for potential differences in recognizability due to individual differences in drawing skill and systematic differences in the amount of detail in drawings across conditions, we next considered an augmented model containing two additional covariates: self-reported drawing skill and the number of strokes in each drawing. We found that this augmented model outperformed the original one without these covariates by a large margin ($\chi^2 = 62.3$, $p = 2.9\text{e}{-14}$), suggesting that these additional variables account for meaningful additional variance in drawing categorization.

Further examination of the model coefficients revealed that exemplar drawings were less categorizable than category drawings (exemplar: 0.802, 95% CI: [0.739, 0.854], category: 0.847, 95% CI: [0.794, 0.888]; $b = -0.48$, $z = -4.88$, $p = 1.1\text{e}{-6}$), photo-cued drawings were less categorizable than label-cued drawings (photo: 0.797, 95% CI: [0.732, 0.849]; label: 0.852, 95% CI: [0.800, 0.892]; $b = 0.22$, $z = -2.29$, $p = 0.026$), and the gap between exemplar and category drawings was more pronounced in the photo-cue condition ($b = 0.33$, $z = 2.36$, $p = 0.018$; Fig. 1.3B). These results show that communicative goal and cue type interact to impact the amount of category evidence in a drawing. Specifically, drawings intended to depict a specific exemplar that was just seen contain less category-diagnostic information than drawings intended to evoke the general category. However, when participants are prompted with only a category label, the effect of target communicative goal is much more modest.

The finding that photo-cued drawings were less recognizable at the category level merited further investigation given a previous study that had varied cue type had found the opposite pattern of results (J. Fan et al., 2018), with photo-cued drawings outperforming label-cued ones. That finding had been interpreted as potentially reflecting the ability of photos to remind participants of category-diagnostic visual details that may have otherwise been difficult to retrieve from long-term knowledge. Towards reconciling these two sets of findings, we explored the possibility that the degree to which photo cues impact a drawing's categorizability depends on how prototypical the photographed exemplar is. Specifically, less typical photos may yield less categorizable drawings because these photos contain less category-diagnostic information for participants to draw upon. To evaluate this possibility, we fit categorization judgments for photo-cued drawings with a mixed-effects logistic regression model containing both cue typicality, communicative goal, and their interaction as fixed effects, and random intercepts for drawing participant, categorization participant, and target category. We found that this model outperformed reduced variants lacking the interaction term ($\chi^2 = 18.5$, $p = 1.7\text{e}{-5}$), communicative goal ($\chi^2 = 17.6$, $p = 2.7\text{e}{-5}$), and typicality ($\chi^2 = 129.4$, $p < 2\text{e}{-16}$) as predictors, suggesting that all three terms explained meaningful amounts of variation in categorization performance. Inspection of the coefficients of this model revealed that, indeed, less typical photo cues yielded less categorizable drawings ($b = -0.25$, $z = 4.62$, $p = 3.9\text{e}{-6}$), with the effect of typicality being greater for exemplar drawings ($b = 0.31$, $z = 4.34$, $p = 1.44\text{e}{-4}$; Fig. 1.3C). Together, these findings suggest a more nuanced view of when photo cues help people produce drawings that are easier to categorize, and when they do not.

How does communicative goal and cue type impact exemplar-level identification?

In the prior analysis, we found that participants intending to depict an exemplar, especially a less typical one, produce drawings that are less recognizable at the category level. One possible explanation for these results is that the task of faithfully capturing the visual appearance of these exemplars was sufficiently challenging that participants simply failed to encode meaningful semantic information in their drawings at all, whether at the exemplar or category levels. Under this account, exemplar drawings would also underperform on the exemplar identification task. Alternatively, these drawings may have succeeded in encoding the distinctive visual details that support recognition at the exemplar level, leading them to outperform category drawings on this recognition task.

To tease apart these possibilities, we constructed mixed-effects logistic regression models with the same random effects structure as above to analyze how easily photo-cued drawings could be matched with their corresponding photo cue. We found using nested model comparison that communicative goal had a large impact on exemplar identification ($\chi^2 = 158.4$, $p < 2\text{e}{-16}$), with exemplar drawings achieving reasonably high accuracy and substantially outperforming category drawings (exemplar: 0.572, 95% CI: [0.532, 0.611]; category: 0.204, 95% CI: [0.179, 0.231]; $b = 1.65$, $z = 15.4$, $p < 2\text{e}{-16}$; Fig. 1.3E). When we augmented this model with information about cue typicality, we discovered that typicality had an impact on exemplar identification, but in different ways depending on communicative goal ($b = 0.248$, $z = 3.53$, $p = 4.2\text{e}{-4}$; Fig. 1.3F). Specifically, while category drawings were more easily matched to their corresponding photo cue when the cue was more typical, exemplar drawings were more difficult to match to more typical photo cues. This pattern of results may reflect the shifting demands of the exemplar-recognition task depending on the typicality of the photo cue: for category drawings, this task is more like a category recognition task when the cue is more typical; for exemplar drawings, this task may be more challenging for more typical photo cues because there are fewer distinguishing features that the drawing and photo can share that are not also shared by the distractor images.

1.7.2 How are exemplar and category recognition related?

To further examine this potential dissociation between category and exemplar information in sketches, we next measured the relationship between variation in the categorization accuracy achieved by an *individual photo-cued drawing* and its exemplar-level identification accuracy (Fig. 1.4). We noticed that individual category drawings that achieved the highest levels of categorization accuracy (i.e., above 75% correct) were also the most difficult to match to their corresponding photo cue (i.e., below 25% correct). On the other hand, while many exemplar drawings still achieved relatively high categorization accuracy, among those that were the most difficult to categorize (i.e., below 25% correct), a relatively large proportion achieved higher identification accuracy. Motivated by these observations, we formally evaluated evidence for a trade-off by constructing a linear mixed-effects regression model predicting identification accuracy for individual drawings from categorization accuracy, goal, and their interaction, with random intercepts for drawing-production participant and target category. This analysis revealed evidence of a

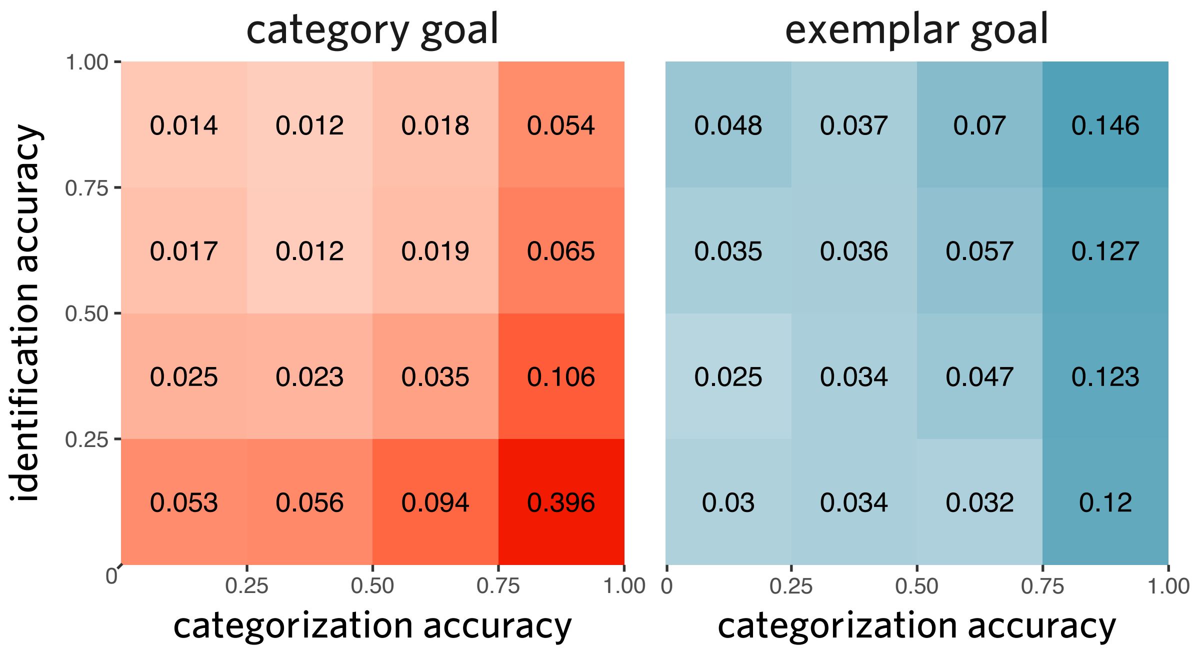


Figure 1.4. Joint frequency distribution of photo-cued drawings achieving different levels of categorization and exemplar identification accuracy under the category goal (left) and exemplar goal (right).

reliable negative relationship between categorization accuracy and identification accuracy ($b = -0.0695$, $t = -3.40$, $p = 6.9\mathrm{e}{-4}$) that did not depend on goal condition. Taken together, these additional findings provide additional support for the notion that category-diagnostic and exemplar-diagnostic information can compete with one another to support drawing recognition at different levels of abstraction.

1.7.3 Sketch feature evaluation

Our results so far support the hypothesis that different visual information is prioritized when communicating about visual objects at different levels of semantic abstraction. When participants have the communicative goal of producing a drawing that faithfully preserves the visual appearance of a specific exemplar, the resulting drawings are less categorizable than drawings intended to evoke a general category, while being easier to match to the specific object. While these patterns point to a tradeoff between more specific and more general information about an object, it remains unclear what features of these sketches may be driving these differences in recognizability beyond mere effort-based measures.

To more directly examine the *semantic features* represented in our sketch dataset, we developed an evaluation criteria involving three distinct measurements: (1) *part complexity* measuring the inclusion of unique object parts, (2) *part emphasis* measuring the number of strokes allocated to each part, and (3) *part spatial arrangement* measuring the spatial layout and compositionality of those parts (Fig. 1.5). To acquire estimates of the part complexity and part emphasis measurements, we gathered dense semantic annotations of every stroke in each drawing of our dataset in order to assess what object parts people included in their drawings. We leveraged VGG-19 and CLIP classifications to evaluate how what the drawn parts looked like and their arrangement might impact on downstream recognizability by naive viewers.

1.8 Method

Building on prior semantic annotation paradigms used to tag individual object parts in drawings from large-scale datasets (Long et al., 2024), we first crowdsourced part decompositions of each object category in order to generate rich sets of object part labels and then crowdsourced part tags of how each stroke in every drawing corresponded to object parts of their target category.

1.8.1 Part decomposition task

1.8.2 Participants

150 participants (71 female, 26.15 years) were recruited from Prolific and completed the study. We did not exclude data from any participants, as none met our exclusion criteria.

1.8.3 Stimuli & Task Procedure

Using the same 32 object categories from Experiment 1, we developed adapted a web-based crowdsourcing platform from prior work to collect object part labels (*Huey et al., 2022). On each trial, participants were cued with a text label of an object category and asked to list 3 to 10 object parts that came to mind (e.g, tail, eye, head for the object concept of "cat"). In one condition, we asked participants to list text labels that best fit their general idea of an object concept such as "cat". In another condition, participants were asked to list text labels that best fit a specific exemplar such as a specific "cat" that came to mind. Participants were instructed to write visually concrete parts of an object (e.g., nouns like "tail") rather than abstract attributes (e.g., adjectives like "fluffy"), to use commonly known names rather than technical jargon (e.g., "stifle"), and to make a complete list of parts for each object category.

1.8.4 Data preprocessing

We applied lemmatization to the resulting part lists to remove syntactically redundant labels (e.g., "paw" vs. "paws"). We also manually edited part labels that were spelled incorrectly or semantically redundant for the object category (e.g., "fur" vs. "hair"). We selected the top 10% most frequently listed part labels of each condition, which provided us a total of 340 part labels from range of 7-17 possible parts per object category.

1.8.5 Part annotation task

1.8.6 Participants

6,486 participants (3,913 female, 26.4 years) were recruited from both Prolific (N=2,105) and our university study pool (N=4,381) and completed the study. We excluded data from 553 additional participants for experiencing technical difficulties with the web interface (N=46) and for having low accuracy on our attention-check trial (N=507). Data collection stopped when every drawing had received annotations from at least three

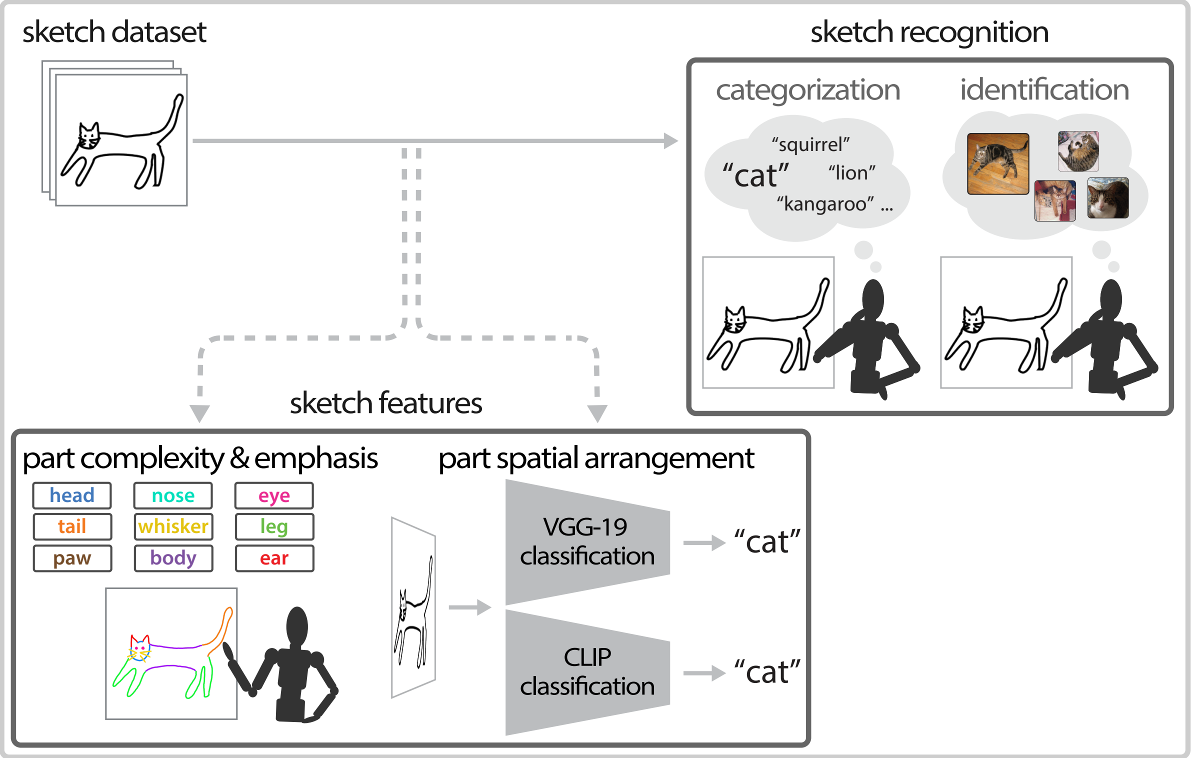


Figure 1.5. Overview of sketch feature analyses and sketch recognition tasks

annotators.

1.8.7 Stimuli & Task Procedure

Leveraging the labels generated from our part decompositions task, we then aimed to systematically measure the semantic information depicted in each drawing. To accomplish this, we designed a second web-based crowdsourcing paradigm using similar tools used in previous research (Huey et al., 2023; Mukherjee et al., 2019) to assign a part label to each pen stroke of each drawing in the collected dataset. On each trial, annotators were presented with a drawing, name of the corresponding object category (e.g., "cat"), and gallery of the part labels associated with object category. For each stroke in the presented drawing, annotators were prompted to tag it with the part label that best described that part of the object that it represented. If a stroke appeared to represent multiple different parts, they were allowed to tag it with multiple part labels. Additionally, if annotators believed that none of the provided part labels could adequately describe a stroke, they could write their own custom label. Strokes could also be labeled as “unintelligible” if a stroke could not be identified (e.g., scribbles, random dots).

Each annotator was presented with a series of 8 drawings randomly sampled from the drawing dataset but consistent broad animate and inanimate object categories, as well as one “attention-check” trial that was randomly inserted into the series. This “attention-check” trial consisted of a pre-selected drawing that was considered relatively easy to annotate and also annotated by a researcher. If annotators failed to match the researcher’s annotation criteria for this drawing, data sessions from these annotators were excluded from subsequent analysis.

1.8.8 Data preprocessing

We first manually inspected the dataset and excluded 124 drawings that were either fully uninterpretable (e.g., scribbles) or did not pertain to the target category (e.g., person drawn in the “blimp” category, lightning bolts drawn in the “ray” category, or horror character drawn in the “saw” category), resulting in 12,164 remaining drawings. Then, to evaluate how often annotators agreed on what each stroke represented in each drawing, we calculated the inter-rater consistency among annotators. Across drawings, annotators agreed on the same part label for 60.23% of strokes. We retained strokes that were assigned the same part label by at least two of three annotators. Of the 340 available part labels, we found that annotators only used 153 to label the full drawing dataset. Because custom labels were infrequently used, we did not include them in analyses. Strokes that were labeled as “unintelligible” were also not counted as distinct parts in analysis, resulting in a total of 140,227 annotated strokes across all 12,142 drawings.

1.9 Results

1.9.1 How does communicative goal and cue type guide part complexity?

Which object parts

To measure part complexity, we calculated how often people drew distinct object parts in their representations of the object concepts. We constructed a 153-D binary vector representing the presence of each object part in each sketch. We hypothesized that the object parts included in drawings cued by photos would differ from those included in drawings cued by text labels, if different parts are activated in memory depending on whether they were accessed through visual experience or not. Additionally, we hypothesized that drawings intended to convey a specific exemplar information contain unique object parts that would vary from drawings intended to convey more general categorical information. To evaluate this, we computed the metric distances between sketches from pairs of two different drawing conditions. Following the idea of permutation test (with N=1000), we shuffled the drawing conditions across all drawings within each category and computed the metric distances for each possible pair of conditions to construct a new distribution of metric distances under a null hypothesis. We then assessed the significance of observed metric distances against this distribution derived under the null hypothesis and employed Jensen–Shannon divergence to measure the distances between drawing conditions. Our analysis indicated no significant differences in part complexity across communicative goals and cue types (between category-text and exemplar-photo: $p = 0.0533$, between category-photo and exemplar-photo: $p = 0.2092$, between category-photo and exemplar-text: $p = 0.2818$, between category-text and exemplar-text: $p = 0.1967$, between category-text and category-photo: $p = 0.2151$, between exemplar-photo and exemplar-text: $p = 0.1869$).

How many object parts

We next examined whether people drew more or less unique object parts depending on their drawing condition. We hypothesized that drawings aiming to highlight distinctions among exemplars would prioritize object parts differently compared to those focusing on categorical distinctions. To examine this hypothesis, we analyzed the relationship between the number of object parts included in the drawings and the conditions in which these drawings are produced. We fit a linear mixed-effects model to predict the number of parts based on the communicative goal and cue type, with random intercepts for each category. Our analysis showed that this model significantly outperformed the null-hypothesis model that only accounted for random effect for category in predicting the number of parts ($\chi^2 = 146.89$, $p = 2.2e-16$), suggesting a strong correlation between the drawing conditions and the number of parts depicted. Specifically, drawings focused on communicating exemplar-level information included a larger number of parts (exemplar: 6.68 parts, 95% CI: [6.02, 7.34]; category: 6.30 parts, 95% CI: [5.63, 6.96]; $b = 3.822e-1$, $t_{XXX} = 10.97$, $p < 2e-16$), while the type of cue did not significantly affect the number of parts. These results are consistent with and extend our previous finding that people tend to use more strokes when conveying exemplar-level information and that these strokes translate into more object parts, relative to category-level information,

1.9.2 How does communicative goal and cue type guide part emphasis?

To measure the extent to which people may visually emphasize different object parts depending on their drawing condition, we next calculate the proportion of strokes that participants allocated to unique object parts in their drawings, which resulted in a 153-D distribution representing the proportion of strokes for each object part in each sketch (Fig. 1.6). Although we previously found that participants tended to draw the same object parts of target categories across drawing conditions, we hypothesized that participants cued by photos might devote a greater number of strokes to object parts that tend to be visually larger (e.g., elephant body vs. elephant head), relative to when they are cued by text cues. Additionally, we hypothesized that drawings intending to convey exemplar information about a specific object might depict more strokes to distinct parts, compared to drawings conveying general categorical information.

To evaluate these hypotheses, we applied the same distance metric protocol used to evaluate part complexity and again employed Jensen-Shannon divergence to evaluate the proportion of strokes allocated to object parts across drawing conditions. We found that drawings intending to convey exemplar-level information and cued by photos are significantly different from those conveying category-level information and cued by text, in terms of part emphasis ($p = 0.0289$). However, there is no significant difference in part emphasis between other combinations of communicative goals and cue types (between category-photo and exemplar-photo: $p = 0.2075$, between category-photo and exemplar-text: $p = 0.2643$, between category-text and exemplar-text: $p = 0.2725$, between category-text and category-photo: $p = 0.2823$, between exemplar-photo and exemplar-text: $p = 0.1906$).

1.9.3 How does communicative goal and cue type guide part spatial arrangement?

Our previous analyses examined the semantic part content of the drawings, using two metrics of part complexity and part emphasis. Next, we examined to what extent the arrangement of those parts and what those parts looked like differed across drawing conditions. We leveraged VGG-19 and CLIP models for the feature embeddings of the drawings in our dataset and calculated the cosine distance between to measure the difference between drawing conditions. We observed an interaction with cue type in which drawings both intending to convey exemplar-level information and cued by photos are significantly different from the other drawing conditions in terms of both VGG-19

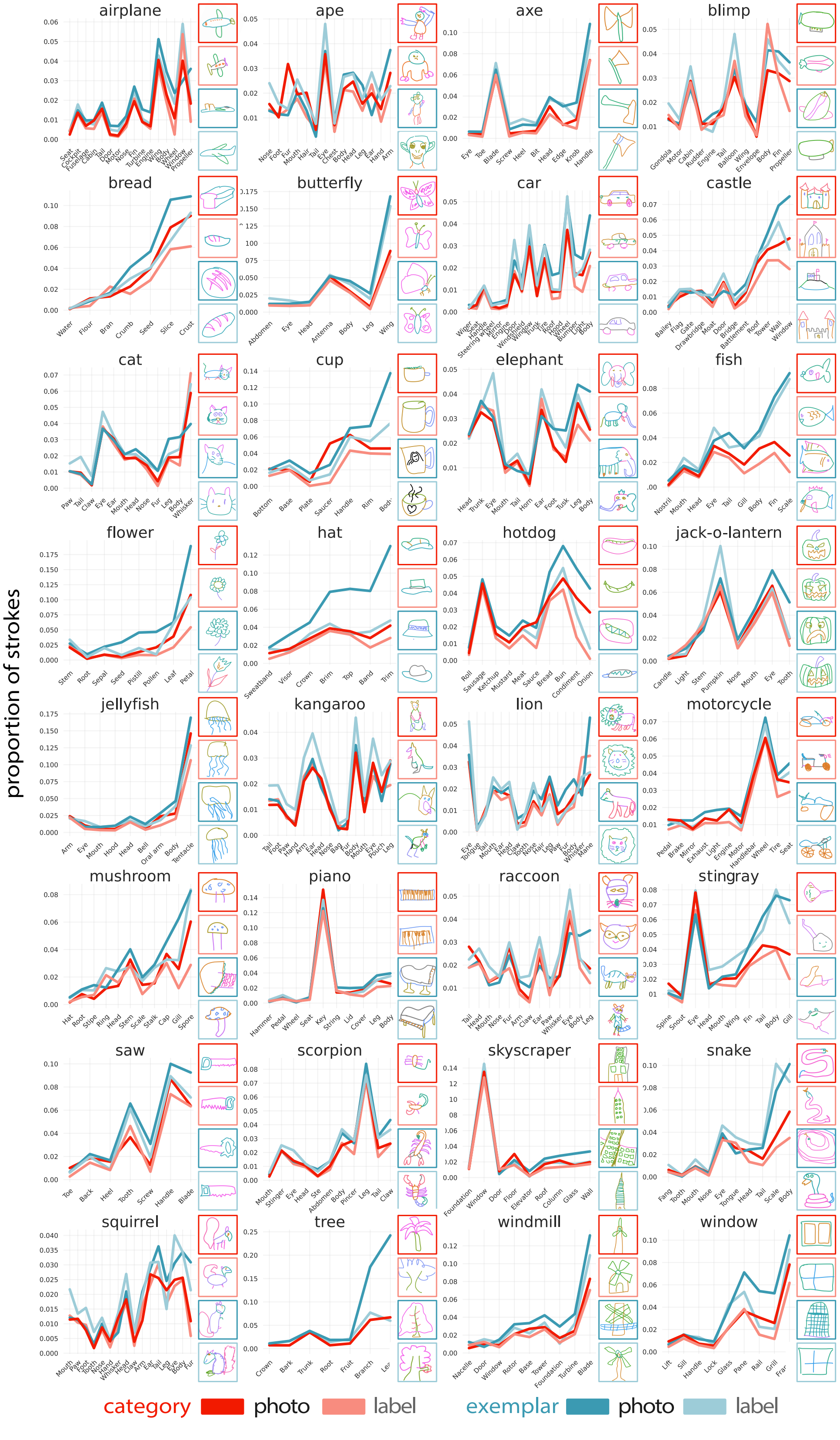


Figure 1.6. Proportion of strokes allocated to object parts across drawing conditions

(between exemplar-photo and category-photo: $p = 0.0174$, between exemplar-photo and category-text: $p = 0.0004$, between exemplar-photo and exemplar-text: $p = 0.0193$) and CLIP (between exemplar-photo and category-photo: $p = 0.0122$, between exemplar-photo and category-text: $p = 0.0028$, between exemplar-photo and exemplar-text: $p = 0.0073$). There is no significant difference between other conditions in VGG-19 and CLIP features.

1.9.4 How do these sketch features relate to recognizability?

Our final goal was assess how well these patterns in sketch features across drawing conditions relate to the differences in the kinds of recognition results observed in Experiment 2. We first investigated how the number of distinct unique parts in a drawing correlated with its recognizability. To do this, we fit a linear mixed-effects model to predict categorization accuracy based on the number of parts, with random intercepts for each category. We also fit linear mixed-effects model to predict identification accuracy, with the same random effects structure as the previous model. We found a significant positive correlation between the number of parts and the accuracy of category recognition ($b = 7.198e-3$, $t_{XX} = 6.662$, $p = 2.81e-11$). The correlation between the number of parts and the accuracy of identification recognition of exemplars was also statistically significant ($b = 5.948e-3$, $t_{XX} = 2.563$, $p = 0.0105$) (Fig. 1.7).

Next, we investigated how all other feature metrics (i.e., part complexity for which features, part emphasis, and part arrangement from VGG-19 and CLIP) correlated with human categorization accuracy. To align these feature metrics with human categorization accuracy, we trained logistic regression classifiers on top of these metrics in a 10-fold stratified cross validation manner to predict the category of each sketch. This provided the categorization accuracy based on part complexity, part emphasis, or part arrangement (VGG-19 and CLIP) features.

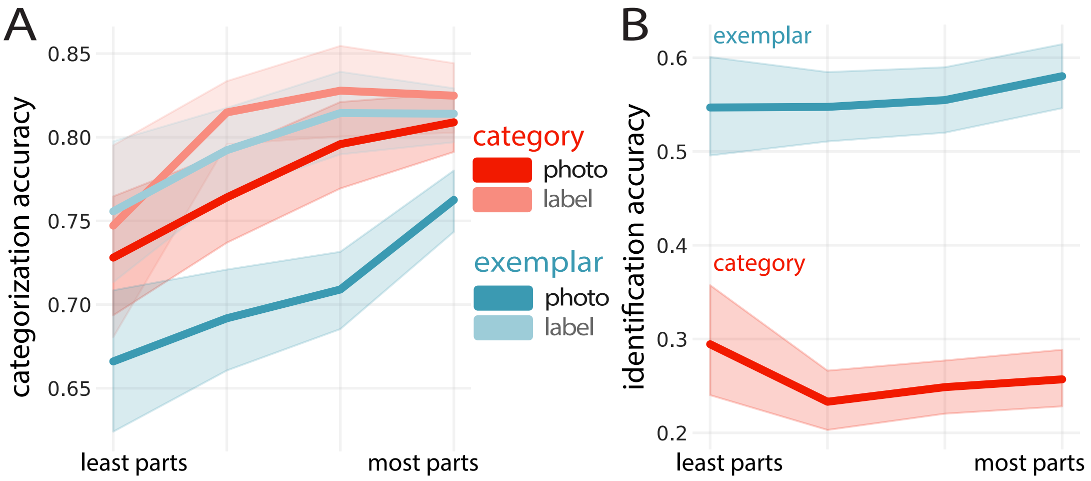


Figure 1.7. (A) Categorization accuracy as a function of quantiled counts of distinct object parts. (B) Identification accuracy as a function of quantiled counts of distinct object parts

1.10 Discussion

How do people choose to externalize their knowledge about visual object concepts when drawing with different communicative goals and when accessing target concepts via visual experience or semantic memory? In the current work, we generated a large-scale dataset of >12K drawings to investigate what semantic information people communicate across their drawings. To systematically explore this, we independently manipulated whether the goal was to produce a drawing of a specific exemplar or a general category, and whether participants had visual access to a photograph of an object or not just before starting to draw. We developed a novel evaluation paradigm to measure an array of part-level information and measured their impact on each drawing's recognizability to naive viewers. We found that exemplar drawings are easier to identify but less categorizable than category drawings, suggesting that people prioritize different diagnostic information in their drawings when drawing visual object concepts at vary levels of abstraction. Moreover, drawings that were cued by photograph are less categorizable than drawings cued by category label, although this gap is reduced when the photograph is of a more typical exemplar. Taken together, our data provide a more nuanced understanding of how drawings encode meaning at different levels of semantic abstraction, suggesting a dissociation between how drawings communicate more general vs. more specific meanings.

A major contribution of our work is publicly offering a densely annotated drawing dataset that spans both a diversity of object categories, as well as a diversity of exemplar level object categories. By systematically measuring the part information in these drawings and as well as systematically manipulating the task contexts under which these drawings were produced, this dataset creates new opportunities to explore the nature of how people map the correspondence between their internal and external representations of visual object concepts. In particular, we believe that future research lines could align with at least three of the following directions.

First, although we interpret differences between the photo-cue and text-cue conditions as being primarily driven by differences in modality (i.e., visual vs. linguistic), these two cue types also varied in other ways— specifically, while photographs inherently capture the appearance of individual exemplars, there are different words to refer to exemplars and to categories at different levels in the semantic hierarchy (Bauer & Just, 2017; E. H. Rosch, 1973). However, in our study we only compared photo cues to basic-level category labels. Thus while we are confident in the differences we measured due to this manipulation, it is not clear to what degree these differences reflect differences in modality, *per se*, as opposed to differences in the level of semantic abstraction. Future work could bring greater clarity to this issue by including a condition in which drawings cued by more specific labels for subordinate categories or even labels for previously seen exemplars were compared with drawings cued by photos.

Second, the present study focused on real-world objects that the participants we recruited would generally be familiar with, allowing them to rely on pre-existing semantic knowledge when producing drawings or recognizing them. As such, the degree to which the dissociation we observed between category-level and exemplar-level information is dependent on such prior knowledge is not clear. In particular, the role of culturally transmitted conventions for depicting visual concepts may be important to consider, in conjunction with the role of direct visual experience with these objects. One promising direction to tease these roles apart in future research would be a replication of the current study with novel object stimuli where participants would not be able to rely upon preexisting semantic knowledge, while still being able to selectively communicate about these objects at different levels of abstraction.

Third, we employed crowdsourcing to obtain empirical measurements of category-level and exemplar-level information in each drawing. A natural follow-up question concerns the precise nature of the computations supporting perception and decision-making in these tasks. Prior work has successfully employed deep convolutional neural networks optimized on categorization tasks to measure category-level information in drawings (J. Fan et al., 2018; J. E. Fan et al., 2019). However, such models may be less well suited to representing fine-grained visual differences between exemplars. Thus another natural direction for future work would be to evaluate a broader array of computational models that embody different hypotheses about the underlying representation and prior experience required to support recognition of drawings at both the category and exemplar levels (J. E. Fan et al., 2020), leveraging more recently developed models optimized for exemplar discrimination (Vinker et al., 2022; Wu et al., 2018; Zhuang et al., 2021).

Taken together, our findings contribute to a growing body of work using drawing behavior to investigate various aspects of cognition, including learning (Chamberlain et al., 2021; J. Fan et al., 2018; Fiorella & Zhang, 2018), communication (J. E. Fan et al., 2020; Garrod et al., 2007; R. X. Hawkins et al., 2019), memory (Bainbridge et al., 2019; Roberts & Wammes, 2020; Wammes et al., 2016), and development (Dillon, 2020; Kellogg, 1969; Long et al., 2024). Such approaches highlight the value of using open-ended production tasks to gain insight into what people perceive and know about the visual world.

1.11 Acknowledgments

I would like to thank my co-authors for their feedback and invaluable contributions to this manuscript and earlier versions of this manuscript: Justin Yang, Xuanchen Lu, and Judith E. Fan. I also thank Bria Long, Mike Frank, and members of the Cognitive Tools Lab at UC San Diego and Stanford University for helpful discussion.

Chapter 1, in full, is currently being prepared for submission for publication of the material. An earlier version of the project was published as Yang, J.& Fan, J. E.. (2021). Visual communication of object concepts at different levels of abstraction. *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*. Cognitive Science Society. The dissertation author was the primary investigator and author of this updated material.

References

Bainbridge, W. A., Hall, E. H., & Baker, C. I. (2019). Drawings of real-world scenes during free recall reveal detailed object and spatial information in memory. *Nature Communications*, *10*(1), 1–13.

Ballester, P., & Araujo, R. (2016). On the performance of googlenet and alexnet applied to sketches. *Proceedings of the AAAI conference on artificial intelligence*, *30*(1).

Barrett, M., & Light, P. (1976). Symbolism and intellectual realism in children's drawings. *British Journal of Educational Psychology*, *46*(2), 198–202.

Bauer, A. J., & Just, M. A. (2017). A brain-based account of "basic-level" concepts. *Neuroimage*, *161*, 196–205.

Bhunia, A. K., Yang, Y., Hospedales, T. M., Xiang, T., & Song, Y.-Z. (2020). Sketch less for more: On-the-fly fine-grained sketch-based image retrieval. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9779–9788.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological review*, *94*(2), 115.

Bozeat, S., Lambon Ralph, M. A., Graham, K. S., Patterson, K., Wilkin, H., Rowland, J., Rogers, T. T., & Hodges, J. R. (2003). A duck with four legs: Investigating the structure of conceptual knowledge using picture drawing in semantic dementia. *Cognitive neuropsychology*, *20*(1), 27–47.

Bremner, J. G., & Moore, S. (1984). Prior visual inspection and object naming: Two factors that enhance hidden feature inclusion in young children's drawings. *British Journal of Developmental Psychology*, *2*(4), 371–376.

Chamberlain, R., Kozbelt, A., Drake, J. E., & Wagemans, J. (2021). Learning to see by learning to draw: A longitudinal analysis of the relationship between representational drawing training and visuospatial skill. *Psychology of Aesthetics, Creativity, and the Arts*, *15*(1), 76.

Degen, J., Hawkins, R. D., Graf, C., Kreiss, E., & Goodman, N. D. (2020). When redundancy is useful: A bayesian approach to “overinformative” referring expressions. *Psychological Review*, *127*(4), 591.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition*, 248–255.

Dey, S., Riba, P., Dutta, A., Llados, J., & Song, Y.-Z. (2019). Doodle to search: Practical zero-shot sketch-based image retrieval. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2179–2188.

Dillon, M. R. (2020). Rooms without walls: Young children draw objects but not layouts. *Journal of Experimental Psychology: General*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Edelman, S., et al. (1999). *Representation and recognition in vision*. MIT press.

Eitz, M., Hays, J., & Alexa, M. (2012). How do humans sketch objects? *ACM Transactions on graphics (TOG)*, *31*(4), 1–10.

Eitz, M., Richter, R., Boubekeur, T., Hildebrand, K., & Alexa, M. (2012). Sketch-based shape retrieval. *ACM Transactions on graphics (TOG)*, *31*(4), 1–10.

Fan, J., Yamins, D., & Turk-Browne, N. (2018). Common object representations for visual production and recognition. *Cognitive Science*, *42*, 2670–2698.

Fan, J. E., Dinculescu, M., & Ha, D. (2019). Collabdraw: An environment for collaborative sketching with an artificial agent. In *Proceedings of the 2019 on creativity and cognition* (pp. 556–561).

Fan, J. E., Hawkins, R. D., Wu, M., & Goodman, N. D. (2020). Pragmatic inference and visual abstraction enable contextual flexibility during visual communication. *Computational Brain & Behavior*, *3*(1), 86–101.

Fan, J. E., Mukherjee, K., Huey, H., Hebart, M. N., & Bainbridge, W. A. (2023). Things-drawings: A large-scale dataset containing human sketches of 1,854 object concepts. *Journal of Vision*, *23*(9), 5975–5975.

Fan, J. E., Yamins, D. L., & Turk-Browne, N. B. (2018). Common object representations for visual production and recognition. *Cognitive science*, *42*(8), 2670–2698.

Fiorella, L., & Zhang, Q. (2018). Drawing boundary conditions for learning by drawing. *Educational Psychology Review*, *30*(3), 1115–1137.

Garrard, P., Lambon Ralph, M. A., Hodges, J. R., & Patterson, K. (2001). Prototypicality, distinctiveness, and intercorrelation: Analyses of the semantic attributes of living and nonliving concepts. *Cognitive neuropsychology*, *18*(2), 125–174.

Garrod, S., Fay, N., Lee, J., Oberlander, J., & MacLeod, T. (2007). Foundations of representation: Where might graphical symbol systems come from? *Cognitive science*, *31*(6), 961–987.

Hawkins, R. D., Sano, M., Goodman, N. D., & Fan, J. E. (2023). Visual resemblance and interaction history jointly constrain pictorial meaning. *Nature Communications*, *14*(1), 2199.

Hawkins, R. X., Sano, M., Goodman, N. D., & Fan, J. W. (2019). Disentangling contributions of visual information and interaction history in the formation of graphical conventions. *CogSci*, 415–421.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Holt, S., Fan, J. E., & Barner, D. (2024). Creating ad hoc graphical representations of number. *Cognition*, *242*, 105665.

*Huey, H., *Long, B., Yang, J., George, K. R., & Fan, J. E. (2022). Developmental changes in the semantic part structure of drawn objects. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*(44).

Huey, H., Lu, X., Walker, C. M., & Fan, J. E. (2023). Visual explanations prioritize functional properties at the expense of visual fidelity. *Cognition*, *236*, 105414.

Jongejan, J., Rowley, H., Kawashima, T., Kim, J., & Fox-Gieg, N. (2017). The quick, draw! dataset.

Kellogg, R. (1969). *Analyzing children's art*. National Press Books Palo Alto, CA.

Kosslyn, S. M. (1976a). Can imagery be distinguished from other forms of internal representation? evidence from studies of information retrieval times. *Memory & Cognition*.

Kosslyn, S. M. (1976b). Using imagery to retrieve semantic information: A developmental study. *Child Development*.

Li, K., Pang, K., Song, J., Song, Y.-Z., Xiang, T., Hospedales, T. M., & Zhang, H. (2018). Universal sketch perceptual grouping. *Proceedings of the european conference on computer vision (ECCV)*, 582–597.

Long, B., Fan, J. E., Huey, H., Chai, Z., & Frank, M. C. (2024). Parallel developmental changes in children's production and recognition of line drawings of visual concepts. *Nature Communications*, *15*(1), 1191.

Lu, X., Wang, X., & Fan, J. E. (2023). Learning dense correspondences between photos and sketches.

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General*, *114*(2), 159.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, *37*(4), 547–559.

Mukherjee, K., Hawkins, R. X., & Fan, J. W. (2019). Communicating semantic part information in drawings. *CogSci*, 2413–2419.

Mukherjee, K., Huey, H., Lu, X., Vinker, Y., Aguina-Kang, R., Shamir, A., & Fan, J. (2024). Seva: Leveraging sketches to evaluate alignment between human and machine visual abstraction. *Advances in Neural Information Processing Systems*, *36*.

Mukherjee, K., Lu, X., Huey, H., Vinker, Y., Aguina-Kang, R., Shamir, A., & Fan, J. E. (2023). Evaluating machine comprehension of sketch meaning at different levels of abstraction. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*(45).

Murphy, G. (2004). *The big book of concepts*. MIT press.

Murphy, G. L., & Smith, E. E. (1982). Basic-level superiority in picture categorization. *Journal of verbal learning and verbal behavior*, *21*(1), 1–20.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of experimental psychology: General*, *115*(1), 39.

Palmeri, T. J., & Gauthier, I. (2004). Visual object understanding. *Nature Reviews Neuroscience*, *5*(4), 291–303.

Potter, M. C. (1976). Short-term conceptual memory for pictures. *Journal of experimental psychology: human learning and memory*, *2*(5), 509.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*, 8748–8763.

Roberts, B. R., & Wammes, J. D. (2020). Drawing and memory: Using visual production to alleviate concreteness effects. *Psychonomic Bulletin & Review*, 1–9.

Rogers, T. T., & Patterson, K. (2007). Object categorization: Reversals and explanations of the basic-level advantage. *Journal of Experimental Psychology: General*, *136*(3), 451.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*(3), 382–439.

Rosch, E. H. (1973). Natural categories. *Cognitive psychology*, *4*(3), 328–350.

Sangkloy, P., Burnell, N., Ham, C., & Hays, J. (2016). The sketchy database: Learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)*, *35*(4), 1–12.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sitton, R., & Light, P. (1992). Drawing to differentiate: Flexibility in young children's human figure drawings. *British Journal of Developmental Psychology*, *10*(1), 25–33.

Song, J., Yu, Q., Song, Y.-Z., Xiang, T., & Hospedales, T. M. (2017). Deep spatial-semantic attention for fine-grained sketch-based image retrieval. *Proceedings of the IEEE international conference on computer vision*, 5551–5560.

Tanaka, J. W. (2001). The entry point of face recognition: Evidence for face expertise. *Journal of Experimental Psychology: General*, *130*(3), 534.

Tanaka, J. W., & Taylor, M. (1991). Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive psychology*, *23*(3), 457–482.

Tversky, B., & Hemenway, K. (1984). Objects, parts, and categories. *Journal of experimental psychology: General*, *113*(2), 169.

Vinker, Y., Pajouheshgar, E., Bo, J. Y., Bachmann, R. C., Bermano, A. H., Cohen-Or, D., Zamir, A., & Shamir, A. (2022). Clipasso: Semantically-aware object sketching. *arXiv preprint arXiv:2202.05822*.

Wammes, J. D., Meade, M. E., & Fernandes, M. A. (2016). The drawing effect: Evidence for reliable and robust memory benefits in free recall. *Quarterly Journal of Experimental Psychology*, *69*(9), 1752–1776.

Wu, Z., Xiong, Y., Yu, S. X., & Lin, D. (2018). Unsupervised feature learning via non-parametric instance discrimination. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3733–3742.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, *111*(23), 8619–8624.

Yang, L., Zhuang, J., Fu, H., Wei, X., Zhou, K., & Zheng, Y. (2021). Sketchgnn: Semantic sketch segmentation with graph neural networks. *ACM Transactions on Graphics (TOG)*, *40*(3), 1–13.

Yu, Q., Liu, F., Song, Y.-Z., Xiang, T., Hospedales, T. M., & Loy, C.-C. (2016). Sketch me that shoe. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 799–807.

Yu, Q., Yang, Y., Liu, F., Song, Y.-Z., Xiang, T., & Hospedales, T. M. (2017). Sketch-a-net: A deep neural network that beats humans. *International journal of computer vision*, *122*, 411–425.

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., & Yamins, D. L. (2021). Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, *118*(3).

Chapter 2

Visual explanations prioritize functional properties at the expense of visual fidelity

Abstract

Visual explanations play an integral role in communicating mechanistic knowledge about how things work. What do people think distinguishes such pictures from those that are intended to convey how things look? To explore this question, we used a drawing paradigm to elicit both visual explanations and depictions of novel machine-like objects, then conducted a detailed analysis of the semantic information conveyed in each drawing. We found that visual explanations placed greater emphasis on parts of the machines that move or interact to produce an effect, while visual depictions emphasized parts that were visually salient, even if they were static. Moreover, we found that these differences in visual emphasis impacted what information naive viewers could extract from these drawings: explanations made it easier to infer which action was needed to operate the machine, but more difficult to identify which machine it represented. Taken together, our findings suggest that people spontaneously prioritize functional information when producing visual explanations but that this strategy may be double-edged, facilitating inferences about physical mechanism at the expense of preserving visual fidelity.

Keywords: natural pedagogy, causal learning, explanation, visual production

2.1 Introduction

From infants exploring the objects in their immediate environment to scientists exploring the frontiers of our solar system, humans are driven to understand how things work and use that knowledge to generate desired outcomes. However, acquiring such mechanistic knowledge from firsthand experience can often be costly in time and effort (Lagnado & Sloman, 2004; Steyvers et al., 2003) and thus the majority of our knowledge about the world depends on its faithful transmission from one generation to another (Boyd et al., 2011; Csibra & Gergely, 2009). This knowledge transmission has long been supported by mechanistic explanations, which help to expose causal relationships latent in otherwise fleeting and complex information (Keil & Lockhart, 2021).

What characterizes good mechanistic explanations, and how do they relate to the phenomena they are intended to explain? Prominent theoretical perspectives highlight several hallmark features (Bechtel, 2011; Wimsatt, 1976), noting that effective mechanistic explanations decompose a causal system into its interacting parts and specify the causal relationships between those parts in the context of a particular function. For example, a bicycle functions by transferring power from the movement of the pedals to the drive wheel via the roller chain between the two wheels, propelling the entire bicycle forward. Such an explanation can be distinguished from a merely descriptive report (e.g., “a bicycle has two wheels, pedals, and a chain), which does not specify the causal relationship between the interacting parts (Corriveau & Kurkul, 2014), and from a teleological explanation (e.g., “a bicycle is for riding from one location to another), which does not decompose the causal system into any constituent parts nor specify how they interact (Kelemen & Rosset, 2009). In addition to playing a key role in scientific theories (Bechtel, 2009), there is growing evidence that mechanistic explanations are also privileged in people’s intuitive understanding of artifacts and biological entities (Chuey et al., 2020; Lockhart et al., 2019). Nevertheless, our understanding of what intuitions people have about what information to prioritize when producing mechanistic explanations themselves is less well developed. Initial insights may be gleaned from prior work investigating the content of explanations that people produce while studying a physical system, which has documented the inclusion of abstract principles (Chi & VanLehn, 1991) and the notion that some explanations may prioritize outward appearance while others emphasize internal properties (Walker et al., 2014; Walker et al., 2017). However, these analyses have generally lacked the resolution to tease apart different hypotheses concerning how people weigh these different kinds of information when constructing a coherent explanation.

While the majority of prior studies investigating explanation behavior have focused on verbal explanations (Chi & VanLehn, 1991; Legare & Lombrozo, 2014; Lombrozo, 2016; Walker et al., 2014; Walker et al., 2017), explanatory *visualizations* may be especially useful for probing the cognitive processes engaged during the communication of mechanistic knowledge (Hegarty, 2011; Mayer, 1999; Scaife & Rogers, 1996; Tversky, 2005). Visualizations of mechanistic phenomena play an important role across scientific domains, including in the biological (Callaway, 2016) and physical sciences (Lipşa et al., 2012). They naturally exploit shape-based and spatial cues to expose both the relevant part-based and relational abstractions that underlie mechanistic understanding (Forbus et al., 2011; Hegarty & Just, 1993; Hegarty et al., 2003; Tversky, 2001), as well as how these abstractions map back onto physical parts of the target system (Bobek & Tversky, 2016; Fan, 2015; Gobert & Clement, 1999; Newcombe, 2013). Moreover, there is ample evidence that visualizations can facilitate learning and inference by comparison with text alone (Glenberg & Langston, 1992; Hegarty & Just, 1993; Larkin & Simon, 1987; Mayer, 1989) by leveraging a small set of relational symbols, such as lines and arrows (Heiser & Tversky, 2006; Tversky, 2005; Tversky et al., 2002; Tversky et al., 2000). However, previous studies that have elicited visual explanations of mechanistic phenomena have not included the detailed analyses of their content that would be required to understand what distinguishes visual explanations in people's minds from other types of visualizations. In particular, while prior work has found that visualizations prompted by functional descriptions of a physical system contain more arrows than those cued by structural ones (Heiser & Tversky, 2006), it remains unclear whether these symbols were simply added to an otherwise ordinary illustration, or whether they formed part of a distinct type of visualization emphasizing information in a substantially different way.

The current studies aim to overcome key limitations of prior work by conducting a thorough investigation of what information people prioritize when generating mechanistic explanations, and leveraging the distinctive properties of *visual* explanations to gain insight into how explanatory abstractions are grounded in our direct experience with mechanical systems. We elicited these visual explanations using an open-ended drawing task, following prior work (Heiser & Tversky, 2006). In Experiment 1, we measure how much people emphasize information about visual appearance or physical mechanisms when producing explanatory drawings of novel mechanical objects, as opposed to depictive illustrations. We used novel objects to probe people's intuitions about how to create informative explanations when generalizing to a specific mechanical system they were not already familiar with, while still being able to rely on prior knowledge about the types of physical mechanisms in play. In Experiment 2, we measure how well naive viewers can map such information back to the corresponding source object. Together, data from these two experiments help to distinguish two potential hypotheses concerning how people generate visual explanations. Under the *cumulative* hypothesis, people first produce a complete depiction of an object's parts, after which they augment this representation with symbols that convey how these parts interact. Under the *dissociable* hypothesis, people intending to communicate mechanistic knowledge refrain from drawing all the parts of the object, instead emphasizing the most relevant ones and how they interact, rather than preserving information about the object's overall appearance. Overall, our results were more consistent with the latter dissociable hypothesis: explanatory drawings emphasized different parts from depictions and more effectively communicated mechanistically relevant information to naive viewers, while less effectively conveying information about an object's visual appearance. Together, these findings suggest that people engaging in visual explanation spontaneously prioritize functional information at the expense of visual fidelity.

2.2 Experiment 1A: Production of visual explanations and depictions

Our first goal was to identify the semantic properties that characterize visual explanations of mechanistic knowledge. To accomplish this, we developed a web-based drawing platform in which participants were presented with a series of novel machines and asked to produce two kinds of drawings: on *explanation* trials, they were prompted to produce visual explanations to help a naive viewer learn how the machine functioned; on *depiction* trials, they were prompted to produce visual depictions to help a naive viewer identify the machine by its appearance. To identify the properties that are distinctive of visual explanations, we use depictions as a baseline for comparison, which were produced in the absence of any explicit goal to communicate causal information about the machines. We chose drawing in our visual production task because it is a basic visualization technique that requires minimal equipment (i.e., any stylus and surface), but is a versatile and accessible technique for communicating information in visual form (Sayim & Cavanagh, 2011). Additionally, people have a robust ability to interpret drawings, despite the fact that drawings produced by novices may omit many details and distort the size and proportion of represented objects (Eitz et al., 2012; Fan et al., 2018). In this experiment, we presented participants with simple machines composed of gears, levers, and pulleys. These parts were chosen since they were likely familiar to participants and are the basic components of more complex compound mechanical systems (Prater, 1994). By using these simple machines, our aim was to gain a purer measure of how people translate their high-level goals of either depicting how a machine functioned or what a machine looked like, without need for expertise in a domain or otherwise extensive familiarity with our task.

2.3 Method

2.3.1 Participants

50 participants (29 male; mean age = 39.1 years) were recruited from Amazon Mechanical Turk for the visual production experiment. Two additional participants were recruited, but their data were not included in the study for not meeting our predefined exclusion criteria (e.g., the drawings consisted of scribbles or were otherwise uninterpretable). In this and all subsequent experiments, participants provided informed consent in accordance with the UC San Diego IRB.

2.3.2 Stimuli

We designed 6 novel machines composed from simple mechanical parts (i.e., gears, levers, pulleys). There were two machines employing each type of part. Half of the mechanical parts in each machine were *causal*, meaning they could be used to produce a desired effect (i.e., turn on a light bulb attached to each machine); the other half of mechanical parts were *non-causal*. To match how visually salient they were, the causal and non-causal parts within each machine were always of the same type (e.g., gear), and were approximately matched in size and number (Fig. 2.2, left). For each machine, we produced a video demonstration of it in which a demonstrator's hand was shown to interact with both the causal and non-causal mechanical parts twice each, in a counterbalanced order, to show that the causal part reliably turned on the light, whereas the non-causal part did not. The order of manipulation was counterbalanced across all machines for a total of 12 video demonstrations. Each video was 30 seconds, and the duration of time in which the researcher manipulated each causal and non-causal part was controlled for through post-production video editing. We also conducted a separate validation study to ensure that participants could generally determine how the machines could be operated to activate the light bulb based on these video demonstrations (see Supplementary Materials).

2.3.3 Procedure

We presented a naive group of participants with a series of 6 videos (one of each machine). After each video finished playing, participants were cued to produce one of two kinds of drawings: on *explanation* trials, they were prompted to produce visual explanations intended to help a naive viewer learn how the machine could be operated to activate a light bulb; on baseline *depiction* trials, they were prompted to produce visual depictions intended to help a naive viewer identify the machine by its appearance (Fig. 2.1). All participants produced three visual explanations and three visual depictions, in a randomized sequence, such that they drew one of each type of drawing for each type of machine. Participants used their cursor to draw in black ink on a digital canvas embedded in their web browser (canvas = 500 x 500px; stroke width = 5px). While drawing on a digital canvas may be more effortful for some participants than drawing on paper, our approach is motivated by prior work that has successfully used digital drawing interfaces to reliably measure variation in drawing production (Bainbridge et al., 2019; Fan et al., 2020; Fan et al., 2018; R. X. Hawkins et al., 2019). Each stroke was rendered in real time on the participant's screen as they drew and could not be deleted once drawn, approximating key aspects of drawing with an ink pen on paper. We reasoned that while it was possible that preventing participants from deleting individual strokes might lead to drawings that sometimes contained extraneous details or strokes produced accidentally, there was no reason to believe that this aspect of the drawing interface would impact one more condition than the other. Participants were not limited in amount of time that they could spend drawing in each trial. At the beginning of each session, participants also completed two practice trials to familiarize themselves with the drawing interface.

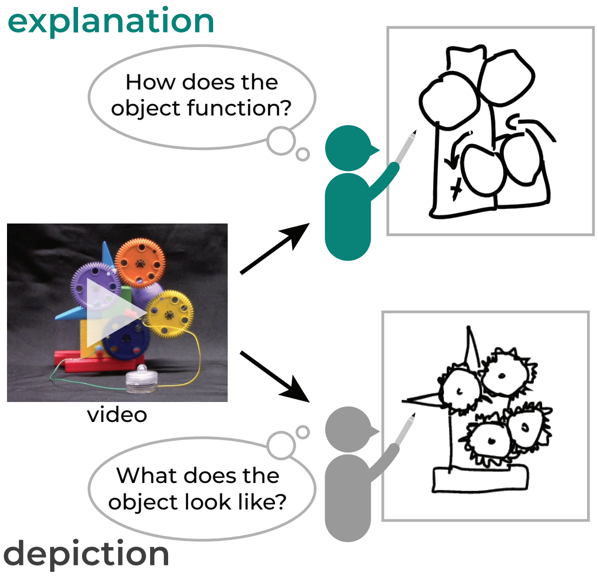

explanation
How does the object function?
video
What does the object look like?
depiction

Figure 2.1. Study 1: Visual Production Task. On each trial, participants viewed a 30-second video demonstrating how to operate a machine to turn on a light bulb. On half of the trials, after the video finished playing, participants were then prompted to produce an explanatory drawing. On the other half of the trials, they were prompted to produce a depictive drawing.

2.4 Results & Discussion

The resulting dataset contained 300 drawings from 50 unique participants: 150 visual explanations and 150 depictions (Fig. 2.2). Insofar as participants are predicted to include more information in visual explanations in accordance with the *cumulative hypothesis*, we predicted that visual explanations would contain more visual detail and take more time to produce, relative to visual depictions. On the other hand, if participants invest a similar amount of effort in both conditions but differ in their semantic content as predicted by the *dissociable hypothesis*, we predicted that the two types of drawings would not substantially differ in how detailed they were nor how much time they took to be produced. To distinguish these possibilities, we analyzed the number of strokes and total drawing time using a linear mixed-effects model predicting the number of strokes from condition and included random intercepts for the type of machine (e.g., gear, lever, pulley) and individual participant.

We found that participants used a similar number of strokes (explanation: 20.33; depiction: 18.9; b = 1.44, t = 1.04, p = 0.301; Fig. 2.1B, *left*) and amount of time drawing in both conditions (explanation: 59300ms; depiction: 57689ms; b = 1144.75, t = 0.359, p = 0.72), suggesting that participants had invested a similar degree of effort when producing both types of drawings. However, while these results provide preliminary evidence against the *cumulative* hypothesis, they also indicate that such simple effort-based measures are insufficient to capture differences in the *semantic information* conveyed by each type of drawing.

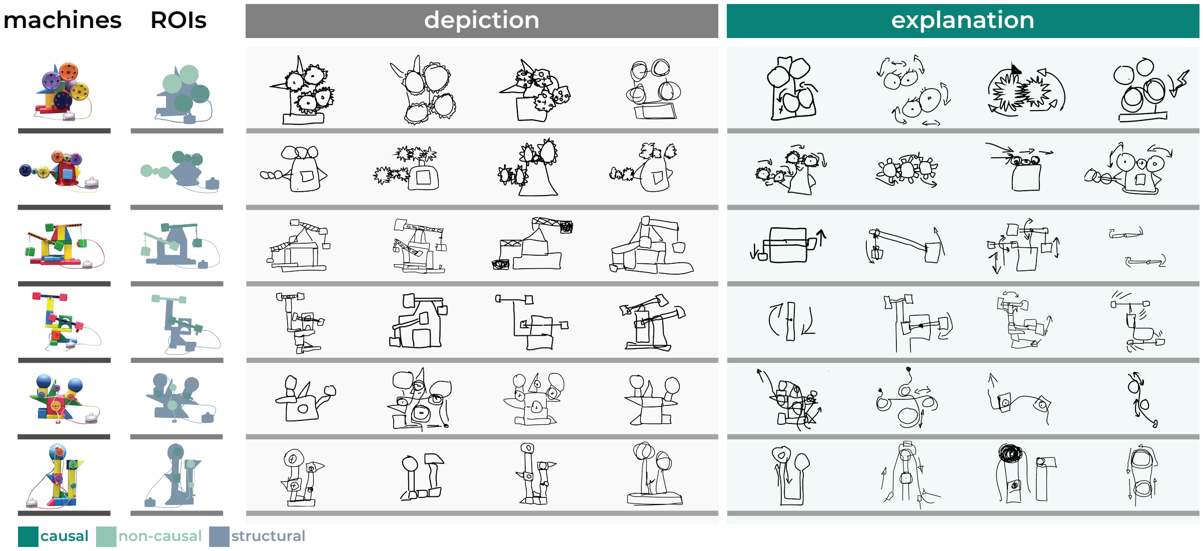


Figure 2.2. Study 1: Visual Production Dataset. ***Left:*** Each machine consisted of multiple mechanical and structural parts. Each region-of-interest (ROI) image indicates the location of both causally relevant and non-causally relevant mechanical parts. ***Right:*** Example depictive and explanatory drawings.

2.5 Experiment 1B: Characterizing semantic content in visual explanations and depictions

To go beyond these effort-based measures, we next crowdsourced annotations from a separate group of naive participant in order to systematically characterize the *semantic information* contained in these drawings. We used these annotations in two ways: first, to understand which parts of the machine participants in Experiment 1A had thought relevant to include in their drawing; and second, to quantify the degree to which each drawing faithfully preserved the relative size and location of each part. One possibility is that visual explanations focus on causally relevant parts, but still faithfully preserve their visual properties. Alternatively, they may distort their visual properties, for example, by making these causally relevant parts more visually salient in their drawing. To distinguish these possibilities, we leveraged techniques from computer vision to precisely measure the differences in the apparent size and location of each drawn part and its actual size and location in the target machine.

2.6 Method

2.6.1 Participants

252 participants (210 male; $M_{\text{age}} = 38.9$ years) were recruited from Amazon Mechanical Turk to provide semantic annotations of the drawings produced in Experiment 1A. We excluded data from 28 additional participants, who did not meet our preregistered inclusion criteria (i.e., low accuracy on attention-check trials, response time <5s).

2.6.2 Task Procedure

Annotators were presented with a set of 10 drawings that were randomly sampled from those drawn in the visual production experiment, as well as reference color photographs of the original machines. In these photographs, each part was color-coded and assigned a unique label and numerical identifier (e.g., 'Gear 2'). Annotators were asked to tag each pen stroke in the drawing based on which part they thought it represented. If a stroke depicted a symbol (e.g., arrow, motion line) rather than a physical part of the machine, annotators were asked to additionally label which part(s) the symbol referred to. If a stroke's meaning was not clear, they could select an "I don't know" option instead. Annotators also completed one attention-check trial that used a drawing from the Experiment 1A dataset that was particularly straightforward to parse and had been manually segmented by the authors. If annotators made 3 or more errors when labeling strokes in the attention-check drawing, all data from that session were excluded from subsequent analysis.

2.6.3 Preprocessing annotation data

For each stroke in every drawing, we obtained labels from at least three annotators indicating which part of the machine it corresponded to (e.g., "gear", "lever", "structural"). Each of these labels were then further grouped into higher-level semantic categories: *causal* strokes representing mechanical parts that were causally related to turning on the light bulb, *non-causal* strokes representing mechanical parts that were not causally related to turning on the light bulb, *structural* strokes representing structural parts, and *symbolic* strokes, including arrows and other marks indicating motion and interactions between parts.

We found that 64.9% of strokes received the same label by all three annotators, and 95.0% of strokes received the same label by at least two of the three annotators. 5.0% of strokes did not reach a majority consensus and received more annotations to resolve this conflict. Moreover, within visual explanations, 55.5% of strokes received the same label by all three annotators, and 93.2% of strokes received the same label by at least two of the three annotators. Within depictions, 75.0% of strokes in depictions received the same label by all three annotators, and 96.9% of strokes received the same label by at least two of the three annotators. In subsequent analyses, we collapsed across annotators and assigned the modal label to strokes which had been given the same label by at least two annotators. For the remaining strokes that did not receive a modal label, we randomly sampled an annotation from the set of annotations that had been assigned to it. We also excluded 5 drawings from subsequent analyses that were deemed to be entirely uninterpretable.

2.6.4 Spatial error analysis

To evaluate how accurately the drawings preserved information about the location and size of each part, we used the following procedure. First, to compute the size and location of drawn parts, we grouped all strokes within a drawing that were tagged with the same semantic label, then determined the coordinates of the rectangular bounding box containing those parts (Fig. 2.3B). For example, if a drawing contained strokes representing four different gears and some structural parts, then this step would yield five bounding boxes, one for each gear, and the fifth containing all structural parts. Strokes representing symbols and/or the light bulb were excluded from analysis. Next, to compute the size and location of target parts, we color-coded each part of the still images of the machines in Adobe Photoshop and grouped all the pixels of the same color. We then calculated the coordinates of the individual bounding boxes for each part. Because the goal of our analysis was to measure how accurately drawings preserved *relative* size and location information, we aligned each drawing to its target machine before computing size and location errors. Specifically, we defined the bounding box containing the entire drawing and the bounding box of the target machine containing the entire machine in the still image, then applied the translation and scaling transformations needed to align these two bounding boxes.

To calculate raw location error for a given part, we computed the Euclidean distance between the centroid of the bounding box for each drawn part and the centroid of the bounding box for the target part. The raw location error for the drawing as a whole was computed by taking the mean of these distances across all parts that appeared in the drawing. We then divided this raw location error by the length of the diagonal of the machine's bounding box to derive a normalized measure of location error, enabling more straightforward aggregation of location error estimates between machines of different sizes. Here, a value of zero indicates that the centroid of a part was drawn exactly in the same location as the centroid of the target part. Additionally, to calculate the raw size error for a given part, we calculated the absolute value of the difference in area between the bounding box of the drawn part and the bounding box for the target part. We then normalized this raw size error by dividing it by the area of the target part, making it easier to aggregate size error estimates between parts of different sizes. Under this procedure, a value of zero indicates that the size of the drawn part exactly matched the size of the target part. And any deviation in size between drawn and target parts increased normalized size error, regardless of whether the drawn part was larger than or smaller than the target part. The normalized size error for a drawing as a whole was computed by taking the mean across all parts that appeared in the drawing.

2.7 Results & Discussion

Insofar as visual explanations place a greater emphasis on functional information than depictions do in accordance with the *dissociable hypothesis*, we hypothesized that visual explanations would contain: (1) more strokes representing causally relevant parts than non-causally relevant parts and (2) more strokes devoted to conveying movement and interactions between parts, such arrows and other symbols, rather than to representing the physical parts themselves. To evaluate the first hypothesis, we constructed a linear mixed-effects model predicting the number of strokes labeled as "causal" from condition and included random intercepts for individual drawing and individual participant. To evaluate the second hypothesis, we constructed a linear mixed-effects model predicting the number of strokes labeled as "symbol" from condition and included random intercepts for the type of machine (e.g., gear, lever, pulley) and individual participants.

Consistent with the first hypothesis, we found that among strokes representing a mechanical part (i.e., gear, lever, or pulley), a greater proportion were devoted to representing causal parts in visual explanations than in depictions (explanation: 58.0%, depiction: 42.0%, $b = 0.382$, $z = 3.44$, $p = 5.9e-4$; Fig. 2.3A). Consistent with the second hypothesis, a higher proportion of strokes in visual explanations were classified as symbols than in depictions (explanation: 24.8%, depiction: 1.0%, $b = 2.48$, $t = 1.39$, $p = 1.67e-1$) and a lower proportion of strokes represented physical parts, including both causal and non-causal parts (explanation: 25.0%, depiction: 45.8%, $b = -2.77$, $t = -4.86$, $p = 1.31e-5$). These results suggest that the goal of communicating mechanistic knowledge leads people to produce drawings that place greater emphasis on causally relevant components and how they move, and less emphasis on static components, even if they are visually salient.

These results are consistent with findings from prior work (Heiser & Tversky, 2006) that has documented an association between drawings explaining how mechanical systems work and the inclusion of arrows. However, this earlier work could not tease apart the degree to which these arrows were simply added to otherwise ordinary depictive drawings ("cumulative" hypothesis), or whether the inclusion of these arrows was accompanied by a general increase in relative emphasis on causally relevant information by comparison with other visually salient, but non-causally relevant information ("dissociable" hypothesis). By collecting detailed semantic annotations of the elements represented in both kinds of drawings, our current findings go beyond prior work to provide direct support for the latter hypothesis.

Insofar as visual explanations exaggerate the appearance of important parts of each machine, we hypothesized that they would not preserve information about their relative sizes and locations as accurately as visual depictions do. Specifically, we predicted

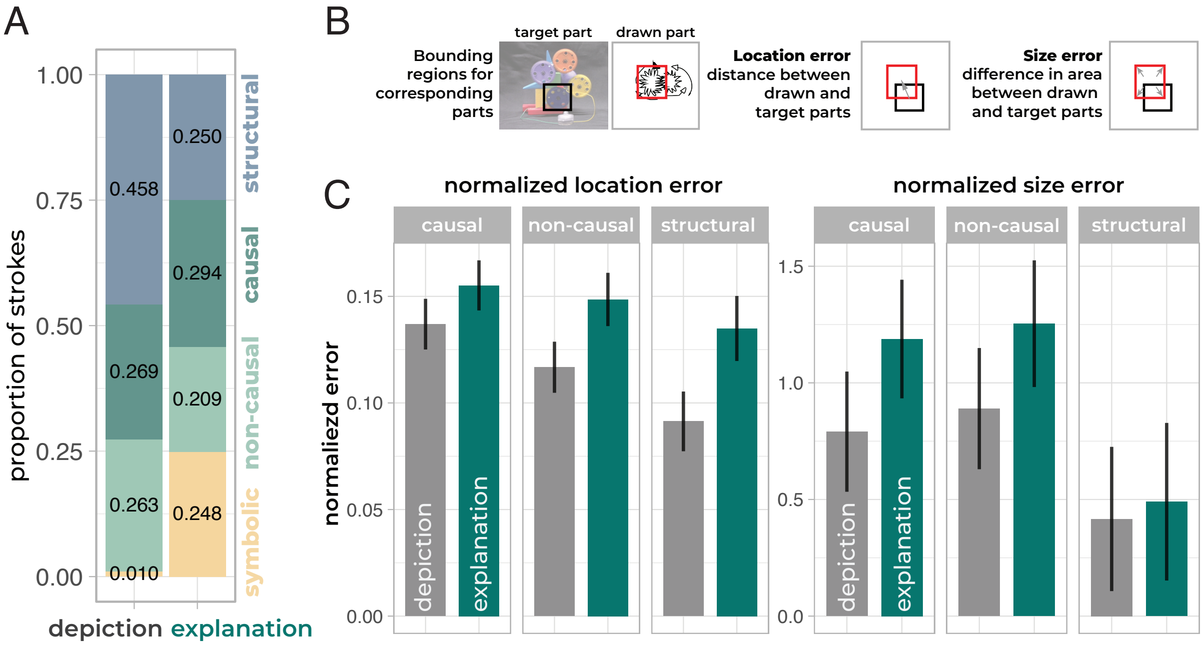


Figure 2.3. Study 1: Results. A: Proportion of strokes conveying different semantic information: *causal* strokes representing mechanical parts that turned the light on; *non-causal* strokes representing mechanical parts that did not; *structural* strokes representing static parts; and *symbolic* strokes, including arrows and other marks indicating motion and interactions between parts. **B:** Accuracy of spatial information in drawings was estimated by defining bounding regions for corresponding parts in each drawing and video, then computing the difference in size and location between the drawn and target parts. **C:** Normalized location and size errors for different semantic part categories. Normalized location errors reflect relative differences between the target and drawn parts, rescaled by the size of the machine. When the normalized location error is zero, the relative locations of each drawn part exactly match the relative locations of each part of the target machine. Normalized size errors reflect relative differences between the target and drawn parts, rescaled by the size of the target part. When the normalized error for size is equal to zero, the relative sizes of each drawn part exactly match the relative sizes of each part of the target machine. Error bars represent 95% CIs.

that: (1) visual explanations might exaggerate the size of causally relevant parts to make them more salient to the viewer and (2) visual explanations might not preserve information about the relative locations of parts, insofar as such information is deemed less relevant for communicating about causal interactions between parts. To evaluate this hypothesis, we fit a linear mixed-effects model predicting the size and location error from condition, including random intercepts of individual machine and participant. We found that mechanical parts were consistently drawn larger in visual explanations than in depictions (explanation: 72.4px, depiction: 60.8px, $b = 8.41$, $t = 1.97$, $p = 4.96e - 2$), in addition to being drawn somewhat further from their actual locations, relative to other parts of the machine (explanation: 75.1px, depiction: 62.3px, $b = 11.6$, $t = 3.15$, $p = 0.18e - 2$; Fig. 2.3C). These findings are consistent with the notion that when explaining how a machine functions, people distort the appearance of functionally relevant parts to make them more salient and discount the importance of preserving exact spatial relationships. Taken together, Experiments 1A and 1B provide evidence that having the goal of communicating mechanistic knowledge systematically affects the kind of information people prioritize when producing visual explanations.

2.8 Experiment 2A: Object identification

However, a critical test of how *useful* such communicative strategies are can be measured how well other people can interpret these drawings to achieve their own behavioral goals. In Experiment 2, we recruited three additional cohorts of naive participants to view the drawings made in the visual production experiment (Experiment 1A) and measured how well each drawing supported their ability to identify the original machine (Experiment 2A), to infer which part of the machine to intervene on to operate it (Experiment 2B), or to infer which action was needed to operate the machine to activate the light (Experiment 2C).

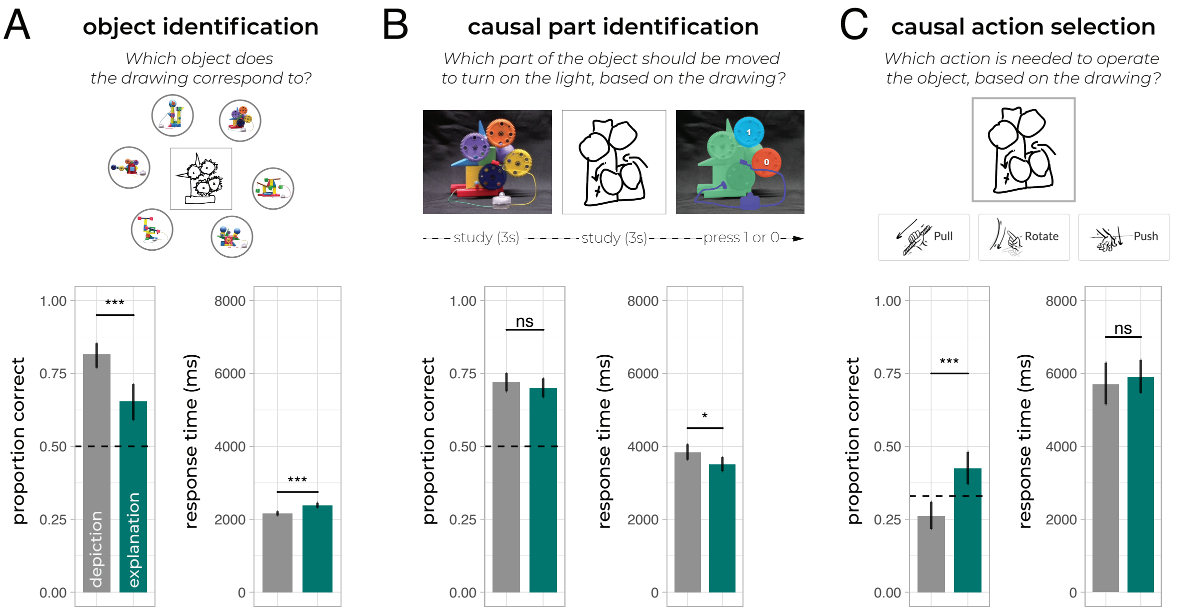


Figure 2.4. Study 2: Visual Inference Tasks and Results. A: In Study 2A, participants identified the machine that matched each drawing. **B:** In Study 2B, participants identified which part of the machine they should intervene on to turned on the light bulb. **C:** In Study 2C, participants inferred which action they would need to perform to turn on the light. Error bars represent 95% CIs.

In Experiment 2A, we hypothesized that the reduced emphasis on structural parts in visual explanations, based on there being relatively fewer strokes devoted to representing them, would make it harder to match it to the original machine, relative to visual depictions. To test this hypothesis, we designed a visual search task to probe how quickly and accurately naive viewers could identify the machine that corresponded to each drawing.

2.9 Method

2.9.1 Participants

50 participants (24 male; $\mathrm{M}_{\mathrm{age}} = 20.5$ years) were recruited from the UC San Diego study pool. Two additional participants were recruited, but data from their sessions were excluded for technical problems (i.e., inability to click on images).

2.9.2 Task Procedure

Each participant was presented with all 300 drawings from Experiment 1A in a randomized sequence. At the beginning of each trial, participants moved their cursor to a crosshair displayed at the center of an empty display. When ready, participants clicked this crosshair to reveal a single drawing (175 x 175px) at that location, surrounded by a circular array of six color photographs (125px x 100px, radius = 250px), one of each machine (Fig. 2.4A). The angular distance between each photo was constant (i.e., 60 degrees) and their angular locations were randomized between trials. Participants were instructed to click on the machine that the drawing corresponded to as quickly and accurately as possible. At the beginning of the session, participants completed 6 practice trials where they were cued with *photos* of each machine (instead of drawings), and had to click on the matching photo in the array.

2.10 Results & Discussion

To investigate how well these drawings support participants' ability to identify the machines, we fit a null model predicting identification accuracy that included random intercepts for different production participants. Although there were 6 machines, we defined chance-level performance at 50%, a theoretical upper bound reflecting our expectation that confusions would be most likely to arise between machines of the same type (e.g., gears).

To evaluate our hypothesis that participants would be slower when presented with visual explanations relative to when they were presented depictions, we fit a linear mixed-effects model predicting response time from condition and random intercepts for individual drawings and participants. Additionally, to evaluate our hypothesis that participants would be less accurate when viewing visual explanations rather than depictions, we fit a mixed-effects logistic regression model to predict individual trial outcomes, with the same random effects structure as our response-time model above.

We found that participants were reliably above chance performance for both types of drawings (explanation: $b = 0.561$, $z = 3.94$, $p = 8.16e-5$; depiction: $b = 1.28$, $z = 10.1$, $p = 2e-16$; Fig. 2.4A). We found that participants were slower to respond (correct trials only: explanation: 2387ms, 95% CI: [2321ms, 2455ms]; depiction: 2161ms, 95% CI: [2103ms, 2220ms]; $b = 9.96e-2$, $t = 5.90$, $p = 1.43e-8$; Fig. 2.4A) and were less accurate when cued with a visual explanation than with a depiction (explanation: 65.4%, 95% CI: [59.0%, 71.0%]; depiction: 81.5%, 95% CI: [77.0%, 85.0%]; $b = -0.847$, $z = -5.033$, $p = 4.84e-7$; Fig. 2.4A, *left*). These results suggest that our manipulation of communicative goals in Experiment 1A measurably impacted how well viewers could extract relevant information from each type of drawing, such that depictive drawings were more informative about the identity of the target machine.

2.11 Experiment 2B: Causal part identification

How well do visual explanations support naive viewers' ability to identify which part of the machines to intervene on to produce desired goals? In Experiment 2B, we hypothesized that greater emphasis on functional parts (i.e., additional strokes, drawn larger), especially those that were causally relevant, would make it easier for learners to infer which component to intervene on to activate the light bulb. To test this hypothesis, we designed another visual search task that probed how quickly and accurately naive viewers could locate the causally relevant part when provided with a drawing of the machine.

2.12 Method

2.12.1 Participants

297 participants (100 male; $\mathrm{M}_{\mathrm{age}} = 28.4$ years) were recruited from Prolific (N=99) and the study participant pool at UCSD (N=198). 8 additional participants were recruited but data from their sessions were excluded, for technical problems with displaying the experimental stimuli (e.g., the videos did not load). We used a larger sample size in Experiment 2B to collect approximately the same number of observations per drawing as we had collected in Experiment 2A.

2.12.2 Task Procedure

Participants were presented with a randomly sampled set of 6 drawings from Experiment 1A, one of each machine, in a randomized sequence. On every trial, participants were presented with three images laid out in a horizontal array, appearing in succession: first, a color photograph of one of the machines appeared on the left; second, after a 3-second delay, a drawing of it appeared in the middle; and third, after another 3-second delay, another photograph of the same machine appeared on the right, this time with one causal part and one non-causal part highlighted in different colors (Fig. 2.4B). Participants were instructed to press a key (i.e., either 0 or 1) to indicate which of the highlighted parts they would intervene on to turn on the light, and to do so as quickly and accurately as possible. At the beginning of the session, participants completed a series of practice trials in which they were familiarized with the task interface.

2.13 Results & Discussion

As in Experiment 2A, we fit a null model predicting identification accuracy that included random intercepts for different production participants to evaluate the degree to which participants performed above chance. To evaluate whether participants would be faster in identifying the causal part when presented with a visual explanation rather than a depiction, we constructed a linear mixed-effects model to predict response time from condition and random intercepts for individual drawings and participants. Additionally, to evaluate our hypothesis that participants would be more accurate for explanations than depictions, we fit a mixed-effects logistic regression model to predict response time from condition and random intercepts for participants. Unlike in Experiment 2A, where each participant saw all drawings produced in Experiment 1A, participants in Experiment 2B were only presented with 6 drawings per session, one of each machine. Given the smaller number of measurements obtained for each drawing in Experiment 2B, we did not have sufficient data to include random intercepts for individual drawings in our statistical models.

We found that both types of drawings supported above-chance performance (explanation: $b = 0.849$, $t = 10.53$, $p = 2e-16$; depiction: $b = 0.919$, $z = 13.04$, $p = 2e-16$; Fig. 2.4B), suggesting that both types of drawings carried meaningful signal about the identity of the causally relevant parts. However, we found that participants were no more or less accurate when cued with a visual explanation than with a depiction (explanation:

70.16%, 95% CI: [67.0%, 73.0%]; depiction: 72.1%, 95% CI: [69.0%, 75.0%]; $b = -9.48e-2$, $z = -0.842$, $p = 0.4$; Fig. 2.4B, *left*). Nevertheless, we did find a small response-time advantage for visual explanations, such that participants were slightly faster to make their response when presented with an explanatory drawing rather than a depictive one (explanation: 3508ms, 95% CI: [3319ms, 3708ms]; depiction: 3840ms, 95% CI: [3635ms, 4057ms]; $b = -9.059e-2$, $t = -2.42$, $p = 1.57e-2$; Fig. 2.4B, *right.*) Taken together, these results do not provide unequivocal evidence that the greater visual emphasis on causal parts in explanatory drawings improved others' ability to more accurately identify these parts *in situ*, although the modest reduction in response time suggests a potential effect on the fluency with which these individuals produced their judgments. Overall, these findings instead suggest that there may be more to the construction of an effective visual explanation than displaying the most functionally important entities more prominently.

2.14 Experiment 2C: Causal action selection

While the prior experiment evaluated how well visual explanations supported naive viewers' ability to identify *where* to intervene on the machines, here we evaluated how well these drawings could support participants' ability to infer *how* to intervene on the machines. In other words, how well do visual explanations support naive viewers' ability to infer which action is needed to successfully operate the machines? Similar to Experiment 2B, we hypothesized in Experiment 2C that greater emphasis on functional parts, especially those that were causally relevant, would make it easier to infer which action was necessary to intervene on the machines to activate the light bulb. To test this hypothesis, we developed a task probing how quickly and accurately naive viewers could identify the appropriate action to perform when provided with a drawing of each machine.

2.15 Method

2.15.1 Participants

267 participants (75 male; $M_{age} = 21.3$ years) were recruited from the UC San Diego study pool. Three additional participants were recruited, but data from their sessions were excluded for technical problems (i.e., videos did not load).

2.15.2 Task Procedure

Participants were presented with a random set of 6 drawings from Experiment 1A, one of each machine, in randomized sequence. On each trial, participants were presented with a single drawing, under which there were 3 buttons labeled ”Pull”, ”Push”, ”Rotate” and ”I don’t know” (Fig. 2.4C). Participants were instructed to click the button that corresponded to the action needed to operate the machine, based on their interpretation of the drawing, and were told to prioritize accuracy. At the beginning on of the session, participants completed a series of practice trials in which they were familiarized with the task interface.

2.16 Results & Discussion

To evaluate the degree to which participants performed the task above chance, we fit a null model predicting accurate responses that was identical in structure to that used in Experiment 2A and 2B. Next, to evaluate differences in how quickly participants could identify the correct action, we fit their responses using the same type of statistical model as in Experiment 2B. Additionally, to evaluate differences in how accurately participants could identify the correct action, we fit their responses using the same type of statistical model as in Experiment 2A and 2B.

We found that participants more accurately identified the correct action when cued with a visual explanation (chance = 33%; explanation: 42.5%, 95% CI: [37.0%, 48.0%];

depiction: 26.09%, 95% CI: [22.0%, 31.0%]; $b = 0.738$, $z = 4.34$, $p = 1.42e-5$; Fig. 2.4C, *left*). Between conditions, they took a similar amount of time to make their response (correct trials only, explanation: 5903ms, 95% CI: [5464ms, 6376ms]; depiction: 5696ms, 95% CI: [5152ms, 6298ms]; $b = 3.56e-2$, $t = 0.555$, $p = 0.579$; Fig. 2.4C, *right*), suggesting that the greater accuracy was unlikely to be due to a speed-accuracy tradeoff. Taken together with the results of Study 2B, these findings suggest that explanatory drawings better supported naive viewers' ability to figure out which action was needed to interact with the machine, even if they did not help them identify which part of the machine to interact with. More broadly, these results show that the visual differences between visual explanations and depictions that we measured in Experiment 1 lead to specific and dissociable consequences on the kind of information people can easily extract from them (e.g., object identity about what the object looks like vs. procedural knowledge about what type of action to use to successfully interact with the object).

2.17 General Discussion

Explanatory visualizations are a crucial tool for conveying mechanistic knowledge, and thus play a key role in many different scientific fields, including biology, physics, and engineering (Callaway, 2016; Chi et al., 1994; Heiser & Tversky, 2006; Lipşa et al., 2012). Nevertheless, there has been a longstanding gap in our understanding of what ordinary people think is relevant when trying to explain how something works, as well as how these visual explanations guide people towards appropriate inferences. Towards closing this gap, here we investigated what information people prioritize when drawing visual explanations of simple mechanical objects (Experiments 1A & 1B). In addition, we measured how well these explanations enabled other people to learn about these objects based on these drawings (Experiments 2A, 2B, & 2C). We found that people spontaneously emphasized functionally important parts of these objects when producing an explanation, using more strokes to draw these parts and making them appear larger than when they only aimed to produce a visually accurate drawing of the object. They also selectively included abstract symbols in their explanations, including arrows and motion lines, suggesting that they believe that providing an explanation means going beyond drawing physical components of the same object. While these explanatory drawings more effectively communicated which action was needed to interact with the object than depictive drawings, this enhancement was accompanied by a loss in diagnostic information about the object's visual appearance. Taken together, our findings suggest that ordinary people can behave in systematic ways when asked to produce a visual explanation, prioritizing information about function (i.e., how parts move and interact) over information about structure (i.e., what parts look like and where they are). This work replicates and extends prior work on visual explanations (Heiser & Tversky, 2006) by showing how they are distinct from other kinds of illustrations not only in terms of what they include (e.g., arrows), but also what they omit (e.g., non-causally relevant parts).

Our findings contribute to a growing body of work characterizing how people evaluate and produce explanatory language (Bobek & Tversky, 2016; Chi et al., 1994; Fiorella & Zhang, 2018; Legare & Lombrozo, 2014; Lombrozo, 2016; Walker et al., 2014; Walker et al., 2017). In this prior work, individuals who are prompted to produce verbal explanations of causal mechanisms also prioritize functional properties over perceptual features that are salient, but not causally relevant (Legare & Lombrozo, 2014; Walker et al., 2014). This pattern of results is broadly consistent with the current study, even though we elicited drawing-based explanations rather than verbal ones. However, our study goes beyond this prior work by further examining how the balance of structural and functional information in visual explanations guide inferences made by downstream learners. We found that visual explanations outperformed visual depictions for supporting some inferences but not others, suggesting that explanations are not necessarily superior to depictions in all settings, but rather a specific tool for conveying knowledge cast at a particular level of abstraction. Moreover, by generalizing prior findings derived from verbal explanations to the visual modality, our work lends support to the notion that similar cognitive mechanisms may account for key aspects of explanatory behavior, regardless of whether these explanations are expressed using words or pictures. Taken together with other recent work extending principles originally developed to account for linguistic phenomena to the visual domain, this body of findings offers converging evidence for a substantial degree of domain generality concerning the mechanisms governing natural communication (Bergen et al., 2016; Fan et al., 2020; Frank & Goodman, 2012).

Our findings also have potential connections to theories of how goals influence how attention is allocated to different elements of a visual scene. In particular, the ability to convey the most goal-relevant information in a drawing may depend not only on what the person producing the drawing is attending to, but also what they expect *someone else* to attend to upon being shown the drawing. Recent work provides some support for the contribution of the former when the goal is to encode the entire visual scene, with visually salient objects being more likely to be included in a drawing (Bainbridge et al., 2019; Harel et al., 2006; Henderson & Hayes, 2017). To what degree does the way that different goals impact how visual attention is deployed across a scene (Chun et al., 2011; Yantis et al., 2000) also determine what information a person is most likely to draw? And how could such influences be differentiated from those providing the basis for adopting the perspective of one's communication partner and thus appropriately emphasizing the information that should be most salient to them (Hawkins et al., 2021), even if it is not what is most salient to oneself? Future studies could investigate the first question by measuring what an individual attends to in a visual scene under different communicative goals, for example by analyzing patterns of eye movements, and relating these measures to which objects they end up including in their drawing. To investigate the second question, future experiments could systematically vary the visual salience of some objects independently of their communicative relevance, which would provide key measurements that could be used to develop and test quantitative theories of how these different factors jointly predict what and how people communicate information in drawings.

Our experimental approach also enables follow-up studies that probe how different kinds of communicative goals may subtly impact the kind of information people believe to be important to include in their explanations. In our study, participants were cued to produce drawings explaining how the machines functioned to produce the desired effect. However, participants may have interpreted these instructions to mean that they should either: (a) explain the specific mechanisms that cause the desired effect for this machine (i.e, how *these* gears turn the light on) or (b) explain the general principles governing the class of mechanisms used by the machine (i.e., how gears work in general). A participant approaching the task with the latter interpretation may be expected to produce drawings that departed more substantially from the visual appearance of the machine than a participant equipped with the former interpretation. Such drawings may be less effective for helping a naive viewer understand any specific machine, but potentially more effective for helping them generalize to a wide variety of machines employing similar physical mechanisms. Future studies could test these predictions directly, shedding light on how the tradeoff between functional and structural information may be modulated by how general a visual explanation is intended to be.

Another key direction for future work is to examine how expertise influences visual explanation behavior. The participants in our studies were unlikely to have received specific training in how to design effective visual explanations, and thus it may not be surprising that the explanations they produced did not outperform depictions in supporting identification of causally relevant parts. One potential explanation for this finding is that, by frequently omitting other (non-causal) mechanical parts and structural parts, these explanations failed to provide enough contextual information to help viewers situate the causally relevant part relative to the rest of the object. Future work could test this hypothesis by prompting drawers to take the perspective of a naive viewer (Shafto et al., 2014), to examine whether they would be more likely to include enough additional structural information to produce more informative visual explanations. Such evaluations may help to clarify the role of perspective taking and pedagogical expertise in the production of explanations that are effective for different audiences.

Overall, this work contributes to our understanding of how visual explanations communicate mechanistic knowledge. In the long run, these studies may lead to both more unified theories of how visual perception, causal reasoning, and social cognition interact to support explanatory behavior, as well as improvements in how visualizations are designed to communicate scientific knowledge in educational and research contexts.

2.18 Acknowledgments

I would like to thank my co-authors for their feedback and invaluable contributions to this manuscript and earlier versions of this manuscript: Xuanchen Lu, Caren M. Walker, and Judith E. Fan. I also wish to thank the members of the Cognitive Tools Lab and Early Learning & Cognition Lab at University of California, San Diego and Stanford University for helpful discussion.

Chapter 2, in full, is a reprint of material as it appears in Huey, H., Lu, X., Walker, C.M., & Fan, J.E. (2023). Explanatory drawings prioritize functional properties at the expense of visual fidelity. *Cognition*. An earlier version of the project was published as Huey, H., Walker, C., & Fan, J. (2021). How do the semantic properties of visual explanations guide causal inference? *Proceedings of the 43rd Annual Conference of the Cognitive Science Society*. Cognitive Science Society. The dissertation author was the primary investigator and author of this material.

References

Bainbridge, W. A., Hall, E. H., & Baker, C. I. (2019). Drawings of real-world scenes during free recall reveal detailed object and spatial information in memory. *Nature Communications*, *10*(1), 1–13.

Bechtel, W. (2009). Constructing a philosophy of science of cognitive science. *Topics in Cognitive Science*, *1*(3), 548–569.

Bechtel, W. (2011). Mechanism and biological explanation. *Philosophy of science*, *78*(4), 533–557.

Bergen, L., Levy, R., & Goodman, N. (2016). Pragmatic reasoning through semantic inference. *Semantics and Pragmatics*, *9*.

Bobek, E., & Tversky, B. (2016). Creating visual explanations improves learning. *Cognitive Research: Principles and Implications*, *1*(1), 27.

Boyd, R., Richerson, P. J., & Henrich, J. (2011). The cultural niche: Why social learning is essential for human adaptation. *Proceedings of the National Academy of Sciences*, *108*(Supplement 2), 10918–10925.

Callaway, E. (2016). The visualizations transforming biology. *Nature*, *535*(7610), 187–188.

Chi, M. T., De Leeuw, N., Chiu, M.-H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, *18*(3), 439–477.

Chi, M. T., & VanLehn, K. A. (1991). The content of physics self-explanations. *The Journal of the Learning Sciences*, *1*(1), 69–105.

Chuey, A., Lockhart, K., Sheskin, M., & Keil, F. (2020). Children and adults selectively generalize mechanistic knowledge. *Cognition*, *199*, 104231.

Chun, M. M., Golomb, J. D., Turk-Browne, N. B., et al. (2011). A taxonomy of external and internal attention. *Annual review of psychology*, *62*(1), 73–101.

Corriveau, K. H., & Kurkul, K. E. (2014). “why does rain fall?”: Children prefer to learn from an informant who uses noncircular explanations. *Child development*, *85*(5), 1827–1835.

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in cognitive sciences*, *13*(4), 148–153.

Eitz, M., Hays, J., & Alexa, M. (2012). How do humans sketch objects? *ACM Transactions on graphics (TOG)*, *31*(4), 1–10.

Fan, J. E. (2015). Drawing to learn: How producing graphical representations enhances scientific thinking. *Translational Issues in Psychological Science*, *1*(2), 170.

Fan, J. E., Hawkins, R. D., Wu, M., & Goodman, N. D. (2020). Pragmatic inference and visual abstraction enable contextual flexibility during visual communication. *Computational Brain & Behavior*, *3*(1), 86–101.

Fan, J. E., Yamins, D. L., & Turk-Browne, N. B. (2018). Common object representations for visual production and recognition. *Cognitive science*, *42*(8), 2670–2698.

Fiorella, L., & Zhang, Q. (2018). Drawing boundary conditions for learning by drawing. *Educational Psychology Review*, *30*(3), 1115–1137.

Forbus, K., Usher, J., Lovett, A., Lockwood, K., & Wetzel, J. (2011). Cogsketch: Sketch understanding for cognitive science research and for education. *Topics in Cognitive Science*, *3*(4), 648–666.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Glenberg, A. M., & Langston, W. E. (1992). Comprehension of illustrated text: Pictures help to build mental models. *Journal of memory and language*, *31*(2), 129–151.

Gobert, J. D., & Clement, J. J. (1999). Effects of student-generated diagrams versus student-generated summaries on conceptual understanding of causal and dynamic knowledge in plate tectonics. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, *36*(1), 39–53.

Harel, J., Koch, C., & Perona, P. (2006). Graph-based visual saliency. *Advances in neural information processing systems*, *19*.

Hawkins, Gweon, H., & Goodman, N. D. (2021). The division of labor in communication: Speakers help listeners account for asymmetries in visual perspective. *Cognitive science*, *45*(3), e12926.

Hawkins, R. X., Sano, M., Goodman, N. D., & Fan, J. W. (2019). Disentangling contributions of visual information and interaction history in the formation of graphical conventions. *CogSci*, 415–421.

Hegarty, M. (2011). The cognitive science of visual-spatial displays: Implications for design. *Topics in cognitive science*, *3*(3), 446–474.

Hegarty, M., & Just, M.-A. (1993). Constructing mental models of machines from text and diagrams. *Journal of memory and language*, *32*(6), 717–742.

Hegarty, M., Kriz, S., & Cate, C. (2003). The roles of mental animations and external animations in understanding mechanical systems. *Cognition and instruction*, *21*(4), 209–249.

Heiser, J., & Tversky, B. (2006). Arrows in comprehending and producing mechanical diagrams. *Cognitive science*, *30*(3), 581–592.

Henderson, J. M., & Hayes, T. R. (2017). Meaning-based guidance of attention in scenes as revealed by meaning maps. *Nature human behaviour*, *1*(10), 743–747.

Keil, F. C., & Lockhart, K. L. (2021). Beyond cause: The development of clockwork cognition. *Current Directions in Psychological Science*, *30*(2), 167–173.

Kelemen, D., & Rosset, E. (2009). The human function compunction: Teleological explanation in adults. *Cognition*, *111*(1), 138–143.

Lagnado, D. A., & Sloman, S. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(4), 856.

Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive science*, *11*(1), 65–100.

Legare, C. H., & Lombrozo, T. (2014). Selective effects of explanation on learning during early childhood. *Journal of Experimental Child Psychology*, *126*, 198–212.

Lipşa, D. R., Laramee, R. S., Cox, S. J., Roberts, J. C., Walker, R., Borkin, M. A., & Pfister, H. (2012). Visualization for the physical sciences. *Computer graphics forum*, *31*(8), 2317–2347.

Lockhart, K. L., Chuey, A., Kerr, S., & Keil, F. C. (2019). The privileged status of knowing mechanistic information: An early epistemic bias. *Child development*, *90*(5), 1772–1788.

Lombrozo, T. (2016). Explanatory preferences shape learning and inference. *Trends in Cognitive Sciences*, *20*(10), 748–759.

Mayer, R. E. (1989). Systematic thinking fostered by illustrations in scientific text. *Journal of educational psychology*, *81*(2), 240.

Mayer, R. E. (1999). Multimedia aids to problem-solving transfer. *International Journal of Educational Research*, *31*(7), 611–623.

Newcombe, N. S. (2013). Seeing relationships: Using spatial thinking to teach science, mathematics, and social studies. *American Educator*, *37*(1), 26.

Prater, E. L. (1994). *Basic machines*. Echo Point Books & Media.

Sayim, B., & Cavanagh, P. (2011). What line drawings reveal about the visual brain. *Frontiers in human neuroscience*, *5*, 118.

Scaife, M., & Rogers, Y. (1996). External cognition: How do graphical representations work? *International journal of human-computer studies*, *45*(2), 185–213.

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cogn. Psychol*, *71*, 55–89.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E.-J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive science*, *27*(3), 453–489.

Tversky, B. (2001). Spatial schemas in depictions. *Spatial schemas and abstract thought*, *79*, 111.

Tversky, B. (2005). Prolegomenon to scientific visualizations. In *Visualization in science education* (pp. 29–42). Springer.

Tversky, B., Morrison, J. B., & Betrancourt, M. (2002). Animation: Can it facilitate? *International journal of human-computer studies*, *57*(4), 247–262.

Tversky, B., Zacks, J., Lee, P., & Heiser, J. (2000). Lines, blobs, crosses and arrows: Diagrammatic communication with schematic figures. *International conference on theory and application of diagrams*, 221–230.

Walker, C. M., Lombrozo, T., Legare, C. H., & Gopnik, A. (2014). Explaining prompts children to privilege inductively rich properties. *Cognition*, *133*(2), 343–357.

Walker, C. M., Lombrozo, T., Williams, J. J., Rafferty, A. N., & Gopnik, A. (2017). Explaining constrains causal learning in childhood. *Child development*, *88*(1), 229–246.

Wimsatt, W. C. (1976). Reductionism, levels of organization, and the mind-body problem. In *Consciousness and the brain* (pp. 205–267). Springer.

Yantis, S., et al. (2000). Goal-directed and stimulus-driven determinants of attentional control. *Attention and performance*, *18*(Chapter 3), 73–103.

Chapter 3

What makes different data visualizations effective for answering different questions about underlying data patterns?

Abstract

Data visualizations are powerful tools for communicating quantitative information. However, while graph production and comprehension research have largely focused on experts with formal data visualization training, little is known about how the general population—the majority of viewers of graphs—thinks about what makes different data visualizations effective for understanding underlying quantitative patterns in the world. In Experiment 1, we asked participants (N=398) which of eight bar graphs would be most useful for answering a particular question, where all bar graphs were generated from the same data but varied in how the data were arranged. We then tested the degree to which participants' preferences aligned with how well viewers (N=542) could answer questions about those same bar graphs. We found participants could discern between graphs that were informative and uninformative for answering different questions. However, while they were biased towards graphs that were at least minimally informative, their decisions did not necessarily reflect a sensitivity to how different graphs could better support graph comprehension by viewers. To further disentangle if people are sensitive to differences in graph informativity, we conducted Experiment 2 in which we only presented participants with informative graphs but that varied in graph complexity and graph type. Although participants (N=119) were biased towards bar graphs plotting fewer variables, we found that their preferences for different graphs reflected a sensitivity towards those that best helped viewers (N=1752) answer questions. Altogether, our findings suggest that people are sensitive to selecting graphs that can effectively convey critical quantitative patterns to viewers. With data visualization being one of the most computationally complex forms of visualization, these insights help develop more unified theories of visual communication aimed at uncovering how we encode and convey our knowledge of the world to others through visual form.

Keywords: data visualization; graph production; graph comprehension; communication

3.1 Introduction

Data visualizations have been a vital technology in modern history, enabling us to distill large-scale quantitative information and communicate complex ideas to others (Pinker, 1990). From modern computational displays of the predicted community spread of pathogens to hand-penned mappings of London's 1854 cholera outbreak, data visualizations have become a ubiquitous tool for communicating patterns in quantitative data. Their power to do so arises from their ability to simplify complex information into a format that can be readily apprehended in visual form (Bertin, 1983; Card, 1999; Franconeri et al., 2021; Tufte, 1983; Tversky, 2001). Critically, different data visualizations—even when generated from the same underlying dataset—can be used to highlight different kinds of information depending on the communicative context. For example, a single bar plot can be used to aggregate many observations to convey the exact magnitude of their mean, whereas multiple bar plots across several panels might be used to convey variation in this mean across groups within the same dataset.

Our ability to judge what information is most relevant to plot is critical for effective communication using graphs. These judgments are core to data visualization design—so much so that they have motivated an array of practical guidelines for effective visualization design (Ajani et al., 2021; Kelleher & Wagener, 2011; Saket et al., 2018). These guidelines are often informed by our empirical understanding of constraints on human perception and information processing (Cleveland & McGill, 1987; Franconeri et al., 2021; Kosslyn, 1989; L. M. Padilla et al., 2018; Rensink & Baldridge, 2010; Shah & Hoeffner, 2002), as well as individual differences in visualization literacy (Börner et al., 2019; Boy et al., 2014; Lee et al., 2019; Mansoor & Harrison, 2018). However, while constraints on graph *comprehension* are often the target of empirical study, graph *production* has rarely been empirically investigated in non-practitioners (Grammel et al., 2010). Nonetheless, genuine visualization literacy encompasses both capacities: the ability to interpret a graph and the ability to produce an interpretable graph. Furthermore, graph production itself depends on two further competencies: the ability to *generate* graphs and the ability to *evaluate* the degree to which a graph is informative. While the former poses some practical barriers—e.g., computing tools and corresponding technical expertise for plotting data—if that requirement is lifted, then it becomes feasible to investigate the evaluative judgments that are integral to graph production and, therefore, also visualization literacy in the broader population.

Indeed, coordinated investigation of both comprehension and production has long been a cornerstone of the study of linguistic communication (Clark & Hecht, 1983; Pickering & Garrod, 2013). Over the past several years, there have been remarkable advances in our understanding of how communicative goals and context impact the production and interpretation of linguistic utterances (Degen et al., 2020; Franke & Jäger, 2016; Goodman & Frank, 2016; Grice, 1975; Kao et al., 2014). Together, this work has provided converging evidence that a core component of natural language use is the ability to deploy mental models of other people to disambiguate meanings (i.e., during comprehension) and to generate expectations about what will be informative to other people (i.e., during production). More recently, these insights have been successfully extended to explain key aspects of how people produce informative pictorial representations in real-time visual communication tasks (Fan et al., 2020), suggesting that these principles may generalize beyond the domain in which they were originally developed.

In the current work, we aim to evaluate people's sensitive to how communicative goals to convey different data patterns to viewers may shift which data visualizations are more or less effective for viewer graph comprehension. Towards this end, we developed two different tasks which we leveraged across two experiments in the current paper. First, we developed a graph selection task in which people decided which of a set of graphs would be most useful for answering a particular question (e.g., "On average, how much higher are ratings of Drama movies compared to Comedy movies?"), where all graphs were generated

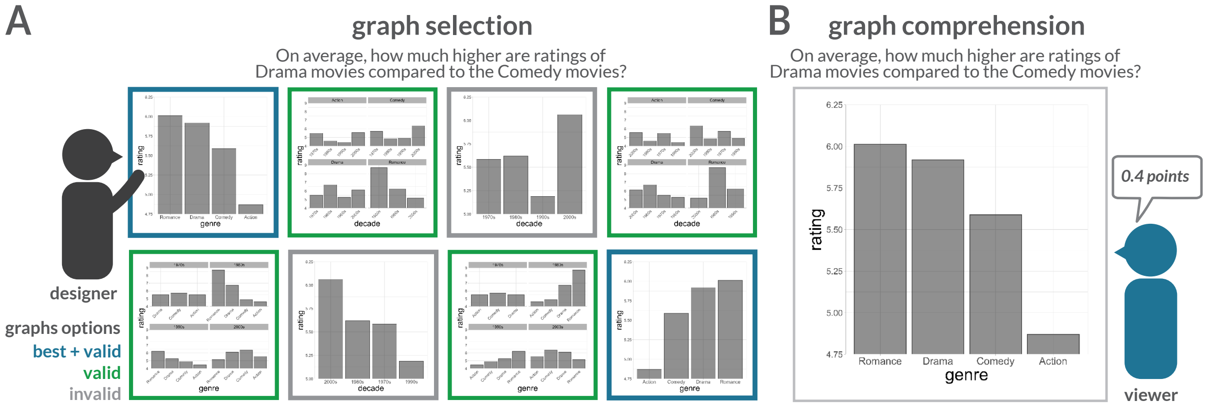
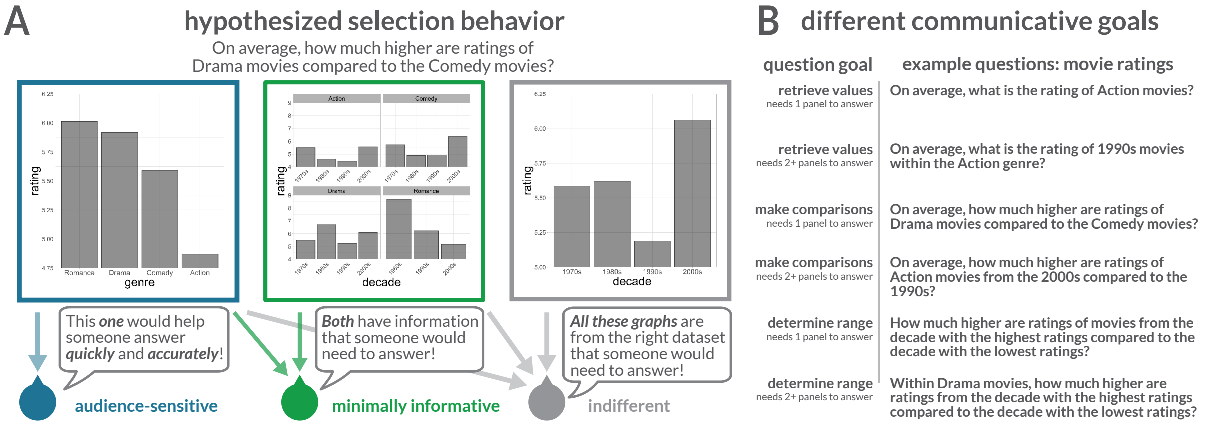


Figure 3.1. Experiment 1: (A) Graph selection task. Participants selected a graph to help a viewer answer an accompanying question as quickly and accurately as possible: *best* graphs are predicted to support fast and accurate comprehension; *informative* graphs contain the minimal information necessary to answer, whereas *uninformative* graphs do not. Graphs are color-coded for this figure, but were not color-coded in the actual task. (B) Graph comprehension task. Viewers answered the same questions using the same corresponding graphs.

from the same dataset but varied in how the data were arranged (see example stimuli in Fig. 3.1A). Second, we developed a graph comprehension task to obtain estimates of how accurately other people ("viewers") could actually answer questions about those data visualizations (Fig. 3.1B). We then examined to the degree to which graphs that best supported viewer graph comprehension were also those that participants in the graph selection experiment were most likely to choose, as well as whether participants' selection behavior varied depending on their prior experience taking more or less formal math courses.

3.2 Experiment 1: How do communicative goals guide beliefs about data visualization effectiveness?

To systematically measure people's preferences for different data visualizations depending on their goal, we developed a graph selection task to measure the range of preferences that people have when trying to communicate specific information to viewers in the form of graphs. Next, to generate behavioral predictions of our audience-sensitive model, we used a graph comprehension task to assess how well naïve viewers could quickly and accurately answer questions about those same graphs. To reduce potential unfamiliarity with different types of data visualizations, we focused on bar graphs, which are one of the most common data visualizations used in education, STEM fields, and journalistic reporting. Additionally, because prior research has suggested that focusing learning on graphing software can lead to student errors (Leonard & Patterson, 2004), we used an alternative-force choice paradigm in our graph selection task in which participants were presented with pre-generated data visualizations.



question goal	example questions: movie ratings
retrieve values needs 1 panel to answer	On average, what is the rating of Action movies?
retrieve values needs 2+ panels to answer	On average, what is the rating of 1990s movies within the Action genre?
make comparisons needs 1 panel to answer	On average, how much higher are ratings of Drama movies compared to the Comedy movies?
make comparisons needs 2+ panels to answer	On average, how much higher are ratings of Action movies from the 2000s compared to the 1990s?
determine range needs 1 panel to answer	How much higher are ratings of movies from the decade with the highest ratings compared to the decade with the lowest ratings?
determine range needs 2+ panels to answer	Within Drama movies, how much higher are ratings from the decade with the highest ratings compared to the decade with the lowest ratings?

Figure 3.2. Experiment 1: (A) Schematic of judgments predicted by hypotheses. (B) Question type examples.

To evaluate how communicative goals guide how people think about what makes data visualizations informative to others, we tested three specific hypotheses. First, if a person's judgments about data visualization design are sensitive to what viewers may need to answer specific questions as quickly and as accurately as possible, we hypothesized that people would prioritize graphs that would help reduce the cognitive effort needed to extract information from them (L. M. Padilla et al., 2018). For example, even if a graph may present all the information necessary to answer a specific question, it may present that information spread across multiple panels and may be mentally laborious to estimate their aggregation if asked to compare information across panels. Therefore, we predicted people would balance two goals: (1) to identify graphs containing the minimal information necessary to answer a presented question (e.g., although a graph may be generated from an appropriate dataset, it may not contain all the information necessary to answer a specific question about it if a specific variable is not plotted); and (2) among those "informative" graphs, to selectively prioritize those that would help viewers quickly and accurately interpret them (Fig. 3.2A, left). We call this the *audience-sensitive* hypothesis. Because data visualizations are more computationally complex and introduced in formal math education in higher grades relative to other forms of visualizations (e.g., drawings), we also predicted that people who have taken more math courses would be more sensitive to the graphs that would better support graph comprehension by viewers.

However, if people are not sensitive to the degree of cognitive effort required by a viewer to interpret a graph, but instead only consider whether a graph contains the minimum information needed to answer specific questions about a graph (i.e., the first goal of the *audience-sensitive* hypothesis), we predicted people would ignore "un-informative" graphs that omit relevant variables but would have uniform preferences among the remaining "informative" graphs (Fig. 3.2A, middle and left). We call this the *minimally-informative* hypothesis.

Lastly, if people's judgments are indifferent to communicative goals and are only concerned about whether a graph is generated from an appropriate dataset, we predicted that they would randomly and uniformly select from *all* the presented graphs (Fig. 3.2A, right, middle, and left). We call this the *indifferent* hypothesis.

3.3 Methods

3.3.1 Participants

398 participants (191 male; $\mathrm{M}_{\mathrm{age}} = 39.6$ years) completed a web-based graph selection task. We excluded data sessions from 3 participants who did not complete the test trials and 7 participants who experienced technical difficulties. Another 542 participants (275 male; $\mathrm{M}_{\mathrm{age}} = 38.4$ years) completed a web-based graph comprehension task. We excluded data sessions from 6 participants who experienced technical difficulties. Participants also completed a post-task survey in which they noted which math courses they had completed by the time of their study participation: "None", "Algebra", "Calculus", and "Statistics". All participants were recruited from Prolific and provided informed consent in accordance with our institution's IRB.

3.3.2 Stimuli

In order to generate a diverse stimuli set of bar graphs, we selected 8 popular datasets from the `MASS` package (Venables & Ripley, 2002). Each dataset contained both numerical and categorical data. We also simplified datasets as needed by filtering out variables, so that the generated graphs would be matched in approximate visual complexity. From each dataset, we generated 8 bar graphs representing means by manipulating three commonly used parameters: (1) grouping in one or multiple separate panels (i.e., faceting), (2) x-axis variable, and (3) organization by ascending ordering of numerical x-axis variables or by alphabetical ordering. Our total test set thus consisted of 64 unique graphs. All graphs were grayscaled so that participants would not be biased by irrelevant aesthetic preferences. Eight additional bar graphs were generated from the `iris` dataset for practice trials.

For each dataset, we generated 6 questions targeting different kinds of information (Fig. 3.2B). Half of the questions asked about information that could only be answered if the correct variables were spread across multiple panels, while the other questions *could* be answered if spread across multiple panels but would be more effective if aggregated into one panel. We also varied which variables were plotted along the x-axis which also determined whether a question was answerable or not. Building on work by Lee et al., 2016 evaluating graph comprehension for different questions, we select three question types to ask participants: to retrieve mean values of a single category; to make comparisons between the means of multiple categories; and to determine the range between the highest and lowest means of categories. The syntax of each question type was standardized across datasets as much as possible. From these questions and the arrangement of variables, bar graphs fell into three categories: (1) "informative" graphs plotting *only* variables necessary for answering questions, which we predicted would be "best" for supporting fast and accurate viewer graph comprehension; (2) 'informative" graphs plotting necessary *and* irrelevant variables; and (3) "uninformative" graphs plotting *only* irrelevant variables that could not be used to answer questions (Fig. 3.1A).

3.3.3 Graph selection task

Participants were presented with a random sequence of 8 trials, each corresponding to a unique dataset. On each trial, they read a description of a dataset and then were presented with a question about the dataset and a 4 × 2 gallery of 8 graphs (Fig. 3.1A). To ensure that participants viewed each graph, they were instructed to use their cursor to hover over each graph, which would then acquire a green border to help participants track which graphs they had "viewed". After viewing each graph, they were instructed to click the one that would best help someone else answer the question as quickly and accurately as possible. Graphs were presented in random order in the gallery, and participants could not select a graph until they had viewed all 8 graphs. The order of presented datasets, as well as the question type corresponding to each dataset, was randomized across participants. Participants also completed one practice trial to ensure that they were familiar with the web interface.

3.3.4 Graph comprehension task

In each trial, participants were presented with a dataset description and corresponding question and provided a numeric answer using a presented graph (Fig. 3.1B). Participants were instructed to answer the question as quickly and accurately as they could, even if they had to guess. In addition to completing one practice trial prior to test trials, participants completed a random sequence of 8 test trials each corresponding to a unique dataset and were not told which graphs were informative or uninformative for a question prompt.

3.4 Results

3.4.1 People are sensitive to differences in informativity between bar graphs of the same dataset

Our main goal was to evaluate people select data visualization to convey different quantitative information. To accomplish this, our first step was to examine whether people select graphs in a non-uniform manner. Using a chi-square goodness-of-fit test, we found that graph selections were non-uniform to different graphs ($\chi^2(7) = 527.13$, $p < 0.001$) and dependent on which question accompanied the graph ($\chi^2(35) = 1590$, $p < 0.001$). These results suggest that participants were sensitive to how graphs vary in informativity for different question prompts. To further explore these results, we then evaluated participants' selections of graphs against a uniform selection predicted by the indifferent hypothesis, which proposes that each graph has the same probability (12.5%) of being chosen. To quantify just how different people's strategies are from these proposed hypotheses, we applied a Jensen-Shannon divergence (JSD) metric. Here, if two distributions perfectly aligned, they would have a JSD of 0. We found that participants' selections were significantly different from the predicted by the indifferent hypothesis (JSD

= 0.51; bootstrapped 95% CI = [0.48, 0.58]; Fig. 3.3), providing evidence that people can reliably discern differences in informativity between graphs generated from the same dataset.

We then evaluated whether having more or less experience with formal math courses may shift how people select different graphs. We conducted a median split on participants based on the number of math classes reported in their post-test survey. We found that the participants who had taken fewer math courses (0-1 courses) and those who had taken more math courses (2-3 courses) selected graphs that were reliably different than the graphs predicted by our indifferent hypothesis (less math: JSD = 0.43; bootstrapped 95% CI = [0.39, 0.53]; more math: JSD = 0.61; bootstrapped 95% CI = [0.56, 0.71]), suggesting that participants even with less math experience are sensitive to differences in informativity between graphs.

3.4.2 People reliably select informative bar graphs over uninformative ones

We next assessed how well participants could help viewers accurately answer questions. We fit a logistic regression predicting the graph type (i.e., informative vs. uninformative) selected with random effects for participant and dataset. Consistent with our minimally informative hypothesis, we found that participants systematically chose informative graphs that contained at least the minimal amount of information needed to answer a corresponding question prompt, relative to uninformative ones ($\hat{\beta} = 2.985$, $z = 15.4$, $p < 0.001$). We found also that this pattern was consistent across participants with previous experience taking more or less formal math courses (less math: $\hat{\beta} = 2.682$, $z = 11.61$, $p < 0.001$; more math: $\hat{\beta} = 3.24$, $z = 12.62$, $p < 0.001$).

To further explore how well participants could discern which graphs were informative or uninformative to answering different questions, we developed a softmax decision rule that prioritized informative graphs ($U = 1$) and discarded uninformative ones ($U = 0$) as predicted by our minimally-informative hypothesis. The softmax temperature was treated as a free parameter for each question. We found that participants' graph selection behavior did not align with the selections predicted by our minimally-informative hypothesis (JSD = 0.12; bootstrapped 95% CI = [0.11, 0.17]; Fig. 3.3), including those with less math experience (JSD = 0.10; bootstrapped 95% CI = [0.08, 0.17]) and more math experience (JSD = 0.18; bootstrapped 95% CI = [0.15, 0.26]). However, we found that participants' selection behavior was better estimated by the selections predicted by our minimally-informative hypothesis, relative to those predicted by our indifferent hypo hypothesis ($\hat{\beta} = -5.628\text{e}{-2}$, $t = -8.469$, $p = 3.77\text{e}{-4}$). These results demonstrate that participants reliably prioritized informative graphs over uninformative ones in order to help viewers accurately answer questions.

3.4.3 People are not necessarily sensitive which informative bar graphs support faster and more accurate graph comprehension by viewers

Our analyses so far reveal that participants prioritize graphs that are informative enough to help viewers answer questions about them. We next evaluated whether people's selection behavior could be explained by a sensitivity in preferences for graph that would support fast and accurate graph comprehension by viewers. Concretely, we hypothesized that if participants were motivated help reduce the cognitive effort needed by viewers to answer questions, we predicted that their selections of graphs would match the graphs that viewers in our graph comprehension task provided faster and more accurate responses. To evaluate this, we first assessed which graphs supported better graph comprehension by viewers. We fit a mixed effect linear regression model to predict viewers' error with random effects for participant and dataset. We fit a second mixed effect linear regression model to predict viewers' response time with random effects for participant and dataset. We found that viewers produced more error when presented with uninformative graphs compared

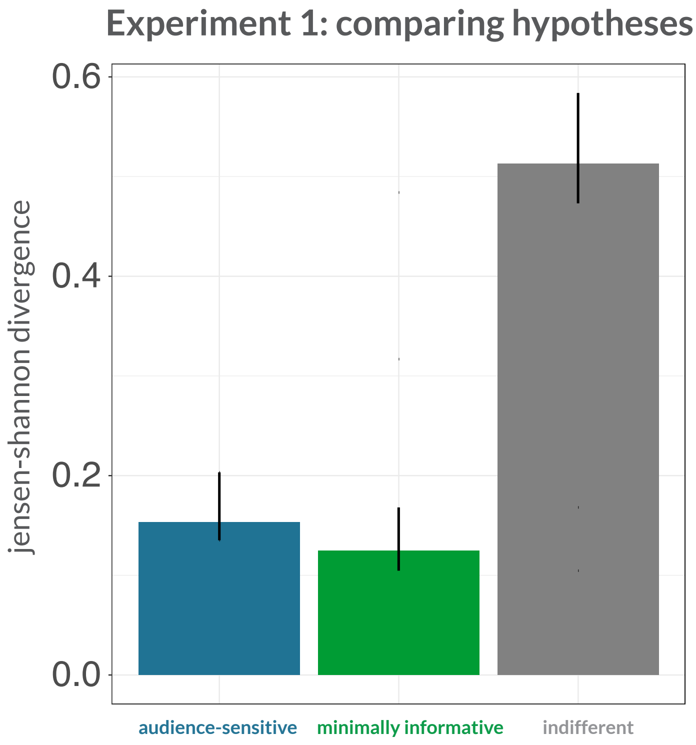


Figure 3.3. Experiment 1: Comparing hypotheses using Jensen-Shannon Divergence against participant selection behavior (identical match = 0). Error bars indicate 95% CIs.

to informative ones ($\hat{\beta} = 35.29$, $t = 10.31$, $p < 0.001$), confirming that informative graphs were more helpful to viewers than uninformative ones. However, we found that viewers responded more quickly when presented with uninformative graphs compared to informative ones ($\hat{\beta} = -2987.36$, $t = -2.011$, $p = 4.44\mathrm{e}{-2}$), suggesting that participants were either faster to respond at the expense of accuracy when presented with uninformative graphs or were faster to identify graphs as uninformative when presented with them. We also applied a likelihood ratio test to a nested model comparison and found that the graph itself explained additional variation in responses ($\chi^2(7) = 37.96$, $p < 0.001$). This analysis provides additional evidence that different graphs were more or less effective for helping viewers answer different question, beyond just whether a graph was informative or not [1].

[1] These results also validate that our stimuli were diverse enough to capture response variation among viewers.

We leveraged these responses by viewers to estimate the graph selections predicted by our audience-sensitive hypothesis. We first calculated the error of viewers' numerical responses, relative to the ground truth answer for each dataset and question type. Aggregating over viewers who received the same graph-question pairing, we computed the root-mean-square-error (RMSE) wherein larger error associated with a graph demonstrates that it did not effectively help answer a specific question. RMSEs were re-scaled between 0 and 1 to normalize across the different data sets. These re-scaled RMSEs were then averaged across datasets and input into a softmax decision rule as utility values. We then used the negative softmax temperature free parameters to estimate which graphs participants would select if they were solely motivated to select graphs that helped viewers answer questions as accurately as possible. These graphs estimated the graph selections predicted by an audience-sensitive hypothesis prioritizing accurate viewers responses. We conducted the same procedure but leveraging the viewers' response times to estimate the graph selections predicted by an audience-sensitive hypothesis prioritizing faster viewers responses.

First, we evaluated whether participants prioritized informative graphs that better supported accurate graph comprehension. To do this, we compared the distribution of graph selections predicted by our audience-sensitive hypothesis against those that participants actually selected (JSD = 0.15 bootstrapped 95% CI = [0.14, 0.20]; Fig. 3.3). Critically, while participants' selection of graphs more closely matched those predicted by our audience-sensitive hypothesis predicting accurate viewer responses compared to the selections predicted by our indifferent hypothesis ($\hat{\beta} = -4.48\text{e}{-2}$, $t = -4.37$, $p = 1.39\text{e}{-3}$), we did not find that participants' selection of graphs more closely matched those predicted by our audience-sensitive hypothesis compared to the graphs selections predicted by our minimally-informative hypothesis ($\hat{\beta} = 9.402\text{e}{-3}$, $t = 1.304$, $p = 0.249$). These results suggest that participants were not necessarily sensitive to which graphs supported more accurate graph comprehension by viewers. We next assessed whether participants who have taken more math courses were more sensitive to the graphs that supported more accurate graph comprehension. We first conducted a split-half reliability test to check the consistency in selection behavior of participants. We found that participants who have taken more math courses were more variable in their selection of graphs (JSD = 0.633) than participants who have taken less math courses (JSD = 0.501). Therefore, although analyses indicate that participants who have taken more math classes did not select graphs predicted by our audience-sensitive hypothesis at a higher rate compared to those who have taken less math courses (less math: JSD = 0.15, bootstrapped 95% CI = [0.14, 0.23]; more math: JSD = 0.19, bootstrapped 95% CI = [0.17, 0.26]), these results are inconclusive because of the variability in graph selections among participants.

Second, we evaluated whether participants prioritized informative graphs that better supported faster graph comprehension. We compared the distribution of graph selections predicted by our audience-sensitive hypothesis against those that participants actually selected (JSD = 0.4 bootstrapped 95% CI = [0.32, 0.48]; Fig. 3.3). We also found that participants' selection of graphs did not significantly from the selections predicted by our indifferent hypothesis ($\hat{\beta} = 2.936\text{e}{-2}$, $t = 0.91$, $p = 0.384$), suggesting that participants did not appear to prioritize graph that supported faster graph comprehension by viewers. Additionally, participants with more math experience were not more sensitive to which graphs would be easier for viewers to make faster responses (JSD = 0.40, bootstrapped 95% CI = [0.38, 0.58]), compared to participants with less math experience (JSD = 0.31, bootstrapped 95% CI = [0.29, 0.45]).

In sum, these results suggest that while people are attentive to which graphs explicitly contain relevant information, they are not necessarily sensitive to subtler differences between how different bar graphs plotting the same data might better support faster and more accurate comprehension by viewers.

3.5 Experiment 2: How do beliefs about graph type and complexity guide beliefs about data visualization effectiveness?

Our current results raise a major question about how specific beliefs about data visualization efficacy constrain people's selection of different graphs. In our graph selection task, we operationalized efficacy as accuracy under time pressures and instructed participants to choose graphs that would help other viewers answer questions as quickly and accurately as possible. We hypothesized that because graphs have a number of unique communicative characteristics that explicitly support large-scale quantitative information compression (e.g., bar graphs represent means), participants would select bar graphs that could most easily and directly help others answer a prompted question. The analyses of Experiment 1 indicated that participants were not necessarily sensitive to graphs that could support faster and more accurate comprehension by others but nonetheless, reliably distinguished informative graphs from uninformative ones. These results suggest that, when presented with informative and uninformative bar graphs, people may be more concerned with identifying which graphs are informative at all, relative to identifying which graphs support *better* graph comprehension by viewers.

To better evaluate people's sensitivity to graphs varying in different levels of informativity, we developed a new set of graph stimuli in Experiment 2 that only included graphs that were at least minimally informative or more to our set of questions. Additionally, we included more variation in our stimuli set in order to further generate more divergent performance among viewers performing the comprehension task and therefore, assess how well people's selections of graphs were in line with what would support faster and more accurate graph comprehension by viewers. Our new stimuli varied by: (1) *graph complexity* varying by the number of plotted variables in each graph; and (2) *graph type* by plotting the same variables across bar graphs, line graphs, and scatter plots. While a large majority of data visualization research has leveraged bar graphs to investigate graph production and graph comprehension, research investigating other graph types have shown that line graphs and scatter plots may be better at helping viewers perceive trends in data, especially in time series data (Wang et al., 2017). Thus, it may be possible that people are sensitive to how different graph types may be more effective for different questions. Consistent with our original hypotheses, our new stimuli offer related but more nuanced hypotheses about what motivations may guide people's judgments about effective data visualization design.

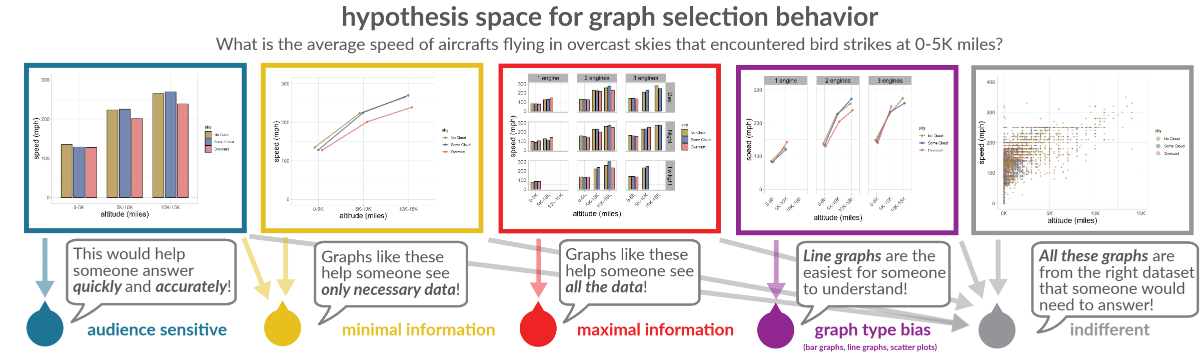

hypothesis space for graph selection behavior
What is the average speed of aircrafts flying in overcast skies that encountered bird strikes at 0-5K miles?
speed (mph)
altitude (miles)
1 engine
2 engines
3 engines
This would help someone answer *quickly* and *accurately*!
Graphs like these help someone see *only necessary data*!
Graphs like these help someone see *all the data*!
Line graphs are the easiest for someone to understand!
All these graphs are from the right dataset that someone would need to answer!
audience sensitive
minimal information
maximal information
graph type bias
(bar graphs, line graphs, scatter plots)
indifferent

Figure 3.4. Experiment 2: Schematic comparison of judgments predicted by each hypothesis.

First, if people prioritize reducing the cognitive effort needed for viewers to answer target questions, we predicted that people would be biased to select graphs that would help viewers answer target questions as quickly and accurately as possible (Fig. 3.4, left). More specifically, because all graphs of this new stimuli set contained the variables relevant for answering prompted questions, we predicted people make a dual judgment between (1) selecting graphs plotting the least variables; and (2) selecting graphs that people believed may be easier for viewers to answer questions about. This hypothesis is most consistent with our *audience-sensitive* hypothesis in Experiment 1. We again formalized this hypothesis by developing a computational model of a graph designer that is more likely to select graphs that lead to better comprehension by a naïve viewer.

By comparison, if people are impartial to the type of graph presentation (e.g., bar graph vs. scatter plot) so long as it is the simplest presentation possible, we predicted that people may be biased to believe that effective data visualization design is synonymous to plotting the least number of variables, regardless of graph type. Here, we predicted people to uniformly select among graph types that plot the least number of variables (Fig. 3.4, middle left). We call this our *minimal data presentation* hypothesis.

On the other hand, if people are biased to believe that effective data visualization design includes more information to signal greater data transparency and scientific rigor, we predicted that people would be select graphs plotting more variables—even at the potential cost of slower and less accurate graph comprehension by naïve viewers. Conversely to our *minimal data presentation* hypothesis, we predicted people to uniformly select among graph types that plot the highest number of variables (Fig. 3.5, middle). We call this our *maximal data presentation* hypothesis.

Alternatively, if people believe that certain graph types are easier for viewers to interpret regardless of the kind of target question, they may be indifferent to the number of plotted variables so long as a graph contains the variables necessary to answer the questions. In this case, people may uniformly select graphs among the same graph type regardless of the number of plotted variables. Specifically, we would predict people to uniformly select graphs among bar graphs, line graphs, or scatter plots depending on which graph type people may believe may be easiest for viewer graph comprehension (Fig. 3.4, middle right). We call this our *graph type bias* hypothesis.

Lastly, if people are merely satisfied by whether or not a graph contains the relevant variables necessary to answer prompted questions, they may not be sensitive to the degree of cognitive effort required by a viewer to interpret a graph. Given that all graphs in our new stimuli set were designed to be “informative”, we predicted that people would uniformly select graphs across all graph types and levels of complexity. Because this hypothesis represents uniform and random choice among the set of presented graphs, this hypothesis most closely resonates with the *indifferent* hypothesis from Experiment 1

(Fig. 3.4, right).

3.6 Methods

3.6.1 Participants

119 participants (40 male; $M_{\text{age}} = 38.87$ years) completed the web-based graph selection task and were recruited via Prolific. We excluded data sessions from 3 participants who did not complete the test trials and 6 participants who experienced technical difficulties. Another 1752 participants (382 male; $M_{\text{age}} = 21.14$ years) completed the web-based graph comprehension task through our university's undergraduate study pool. We excluded data sessions from 9 participants who experienced technical difficulties. As in Experiment 1, participants completed a post-task survey in which they noted the math courses they had completed. In Experiment 2, we increased the possible math courses to include to gain a more granular perspective on participants' math backgrounds: "None", "Algebra 1", "Geometry", "Algebra 2", "Trigonometry", "Precalculus", "Calculus 1", "Calculus 2", "Calculus 3 (Multivariable Calculus)", "Differential Equations", "Linear Algebra", "Probability and Statistics", "Number Theory", "Real Analysis", and "Abstract Algebra". All participants provided informed consent in accordance with our institution's IRB.

3.6.2 Stimuli

We selected four new datasets from the `MASS` package (Venables & Ripley, 2002). We preprocessed datasets to consist of three categories across four variables, so that the generated graphs would be matched in visual complexity. From each datasets, we then generated three bar graphs, three line plots, and three scatter plots, in which each series of graph types consisted a graph that plotted three variables, four variables (with faceting along the x-axis), and five variables (with faceting along both the x-axis and y-axis), respectively. All x-axis variables were the same across the series of graphs for each dataset and were ordered alphabetically. The total test set consisted of 36 unique graphs. An additional nine graphs were generated from the *penguin* dataset for practice trials. We did not include error bars in any graphs. Because these graphs plotted many more variables than in Experiment 1, we colored each x-axis category as red, blue, and yellow to better visually distinguish variables. We maintained a consistent color palette across all graphs to prevent color biases.

For our question prompts, we asked participants to: (1) retrieve mean values of a single category; (2) make comparisons between means of multiple categories; and (3) predict the values of extrapolated values of categories. These question types were selected to evaluate how well the different graph types could be used to answer questions. In particular, to better gauge people's potential preferences for line graphs and scatter plots, we replaced the question type of determining the range between the highest and lowest means of categories with predicting extrapolated values. Additionally, because all graphs were designed to be "informative" to all prompted questions, there was not a need to design questions targeting graphs that required certain aggregations of variables in different panels. Therefore, in Experiment 2, our question stimuli only contained 3 question prompts per dataset. As before, the syntax of each question type was standardized across datasets as much as possible.

3.6.3 Graph selection task

On every trial, participants read a description of a dataset and then were presented with a corresponding question and icons to either see options for bar graphs, line graphs, or scatterplots. After clicking on an icon, they were presented with a gallery of three graphs of the same type but varying in the number of plotted variables. Participants were instructed to carefully inspect all possible graphs before selecting the one they thought could best help someone else answered the prompted question as quickly and as accurately as possible. On each trial, a timer was shown to count down from one minute and 30

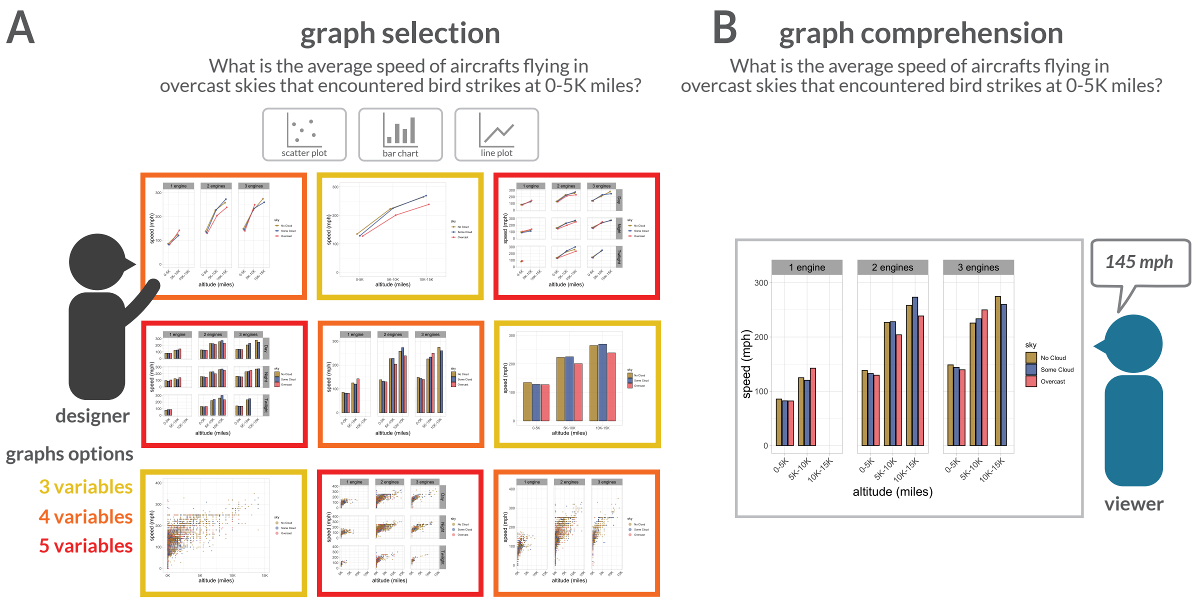


Figure 3.5. Experiment 2: (A) Graph selection task. Participants selected a graph to help a viewer answer an accompanying question as quickly and accurately as possible. Graphs are color-coded for this figure to show levels of graph complexity, but were not color-coded in the actual task. (B) Graph comprehension task. Viewers answered the same questions using the same corresponding graphs.

seconds to encourage participants to make their selections in a timely manner. Graphs were presented in random order within galleries. After participant clicked on their preferred graph, a text box appeared and participants were prompted to write a few words about why they selected a graph.

Each participant was presented with a random sequence of 12 trials, in which four trials were blocked together for a total of three blocks. Within each block, each trial corresponded one of each dataset. Thus, participants were presented with all three question types for each datasets. The order of presented datasets were randomized within block, as well the associated question types were randomized without replacement. Participants also completed three practice trials, which included one of question type, to ensure that they were familiar with the web interface.

3.6.4 Graph comprehension task

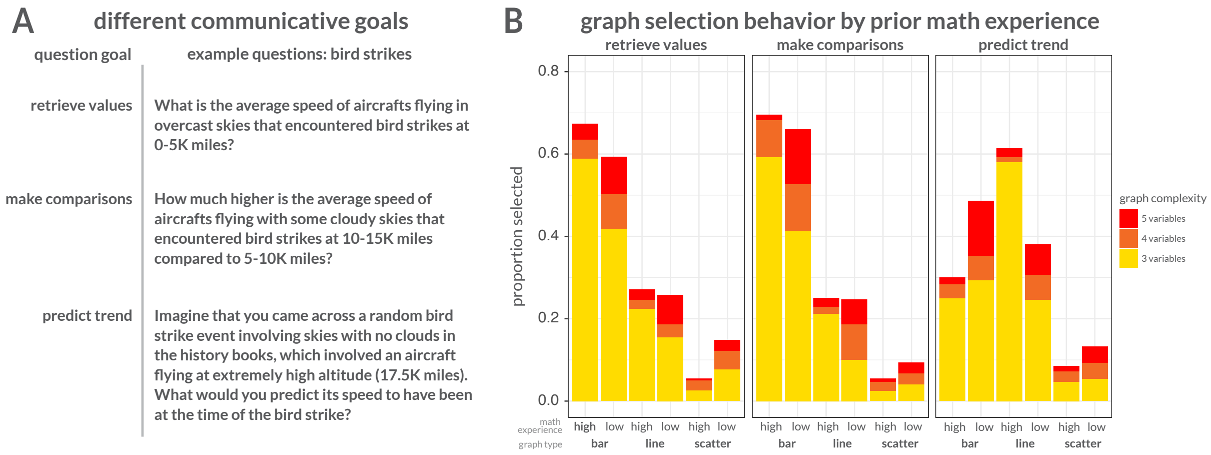
Web design and instructions were identical to the graph comprehension task in Experiment 1.

3.7 Results

3.7.1 People generally prefer towards simpler graphs plotted less variables and bar graphs

We next evaluated people's preferences for different data visualizations in our graph selection task. We first examined whether people selected graphs in a non-uniform manner. Using a chi-square goodness-of-fit test, we found that graph selections were non-uniform to different data visualizations ($\chi^2(8) = 1760.4$, $p < 0.001$) and were dependent on the presented question ($\chi^2(16) = 122.24$, $p < 0.001$). We also compared the distribution of participants' graph selections against a uniform selection distribution as predicted by the indifferent hypothesis in which each graph as the same probability (11.11%) of being chosen. We found that participants' selection behavior was significantly different from our indifferent hypothesis (JSD=0.14; bootstrapped 95% CI=[0.12, 0.18]). These results resonate with those from Experiment 1 and provide further support that people use a richer strategy than randomly selecting among graphs that contain the relevant variables for answering target questions.

To further explore what might explain how people select data visualizations, we assessed participants' preference for different levels of complexity and different graph types. We fit a linear regression predicting frequency of graph selection from graph complexity and found that participants systematically preferred graphs plotting the least number of variables (3 variables: 75.15%, 4 variables: 13.07%, $\hat{\beta} = -90.25$, $t = -6.92$, $p = 7.83\text{e}{-7}$, 5 variables: 11.78%, $\hat{\beta} = -92.13$, $t = -7.059$, $p = 5.76\text{e}{-7}$). We also fit a linear regression predicting frequency of graph selection from graph type and found



question goal	example questions: bird strikes
retrieve values	What is the average speed of aircrafts flying in overcast skies that encountered bird strikes at 0-5K miles?
make comparisons	How much higher is the average speed of aircrafts flying with some cloudy skies that encountered bird strikes at 10-15K miles compared to 5-10K miles?
predict trend	Imagine that you came across a random bird strike event involving skies with no clouds in the history books, which involved an aircraft flying at extremely high altitude (17.5K miles). What would you predict its speed to have been at the time of the bird strike?

Figure 3.6. Experiment 2: (A) Question type examples. (B) Proportion of graphs selected by participants with more or less math experience across graph types and graph complexity.

that participants systematically preferred bar graphs (bar graphs: 56.58%, line graphs: 34.57%, $\hat{\beta} = -32.0$, $t = -4.19$, $p = 4.11\text{e}{-4}$, scatter plots: 8.86%, $\hat{\beta} = -69.38$, $t = -9.087$, $p = 1.01\text{e}{-8}$). However, despite participants' bias toward bar graphs, we found a significant interaction between graph type and question prompt, specifically between bar graphs and line graphs. Specifically, while participants continued to prefer the least complex bar graphs when selecting graphs to help answer questions about retrieving mean values (bar graphs plotting 3 variables: 52.17%, line graphs plotting 3 variables: 19.69%) and comparing means between categories (bar graphs plotting 3 variables: 52.33%, line graphs plotting 3 variables: 16.84%), we found that participants instead preferred line graphs with the fewest variables when selecting graphs to help answer questions about predicting trends (bar graphs plotting 3 variables: 26.68%, line graphs plotting 3 variables: 45.08%, $\hat{\beta} = 14.5$, $t = 4.68$, $p = 1.16\text{e}{-3}$). These results indicate that while participants were biased towards certain bar graphs and simpler graphs, they were also attentive to how different graphs might be better suited to answering different target questions.

We also assessed the degree to which math educational background impacts graph selection behavior relative to graph complexity, graph type, and target question. Conducting a median split on participants based on the number of math classes they reported in their post-test survey revealed, we found that participants with low math experience had taken 0-3 math classes and those with high math experience had taken 4-13 math classes. These results indicate that our sample of Prolific participants, which consists of a more general population than undergraduate study pools, had overall lower math experience relative to our samples of undergraduate students at our university in the Experiment 1 and graph comprehension task in Experiment 2.

Overall, we found that participants of all math backgrounds selected graphs plotting the fewest variables most often (less math, 3 variables: 60.0%; less math, 4 variables: 18.24%, $\hat{\beta} = -47.5$, $t = -11.02$, $p = 1.58\text{e}{-6}$; less math, 5 variables: 21.76%, $\hat{\beta} = -43.5$, $t = -10.1$, $p = 3.31\text{e}{-6}$; more math, 3 variables: 84.89%; more math, 4 variables: 9.75%, $\hat{\beta} = -133$, $t = -96.07$, $p = 7.28\text{e}{-15}$; more math, 5 variables: 5.37%, $\hat{\beta} = -140.75$, $t = --101.67$, $p = 4.37\text{e}{-15}$). However, we found that participants with less math experience selected graphs plotting only 3 variables less frequently than those with more math experience ($\hat{\beta} = -82.0$, $t = -33.13$, $p = 5.03\text{e}{-8}$). Moreover, when participants did select graphs plotting 4 or 5 variables, participants with less math experience tended to do so more often then those with more math experience (4 variables: $\hat{\beta} = 85.5$, $t = 18.89$, $p = 2.57\text{e}{-13}$; 5 variables: $\hat{\beta} = 97.25$, $t = 21.49$, $p = 2.78\text{e}{-14}$). These results suggest participants become more biased towards simpler plots as they gain more formal math training.

Next, we evaluated how preferences for graph types may differ with increased math experience. We found that while participants of all math backgrounds preferred bar graphs (less math, bar graphs: 58.02%; less math, line graphs: 29.45%, $\hat{\beta} = -10.83$, $t = -8.035$, $p = 2.85\text{e}{-9}$; less math, scatter plots: 12.53%, $\hat{\beta} = -17.25$, $t = -12.79$, $p = 2.42\text{e}{-14}$; more math, bar graphs: 55.65%; more math, line graphs: 37.85%, $\hat{\beta} = -31.5$, $t = -7.85$, $p = 2.58\text{e}{-5}$; more math, scatter plots: 6.50%, $\hat{\beta} = -87.0$, $t = -21.68$, $p = 4.46\text{e}{-9}$), participants with less math experience were more biased towards scatter plots than those with more math experience ($\hat{\beta} = 32.25$, $t = 6.54$, $p = 3.84\text{e}{-6}$). Moreover, we found a significant interaction effect between graph type and question across math backgrounds. While participants across both math backgrounds preferred bar graphs with the fewest variables for retrieving values and making comparisons across categories, participants with more math experience preferred line graphs more than bar graphs when predicting data trends (bar graphs: 25.0%, line graphs: 58.05%, $\hat{\beta} = 18.5$, $t = -7.28$, $p = 3.42\text{e}{-4}$), relative to participants with less math experience (bar graphs: 29.33%, line graphs: 24.67%, $\hat{\beta} = -4.0$, $t = -2.86$, $p = 2.89\text{e}{-2}$). These results suggest that as participants with more math experience may become more discerning as to which kinds of graphs are better suited for certain target questions. Simultaneously, these results show that participants are increasingly biased against scatter plots, regardless of graph complexity or target question.

3.7.2 People are sensitive to which graphs support support faster and more accurate graph comprehension by viewers when selecting among informative graphs

Our main goal was to evaluate whether participants were sensitive to the graphs that better supported faster and more accurate graph comprehension by viewers. Using our same procedure as in Experiment 1, we first identified which graphs helped viewers produce answer prompted questions. We fit a mixed effect linear regression model to predict viewers' error with random effects for dataset. We also fit a mixed effect linear regression model to predict viewers' response time with random effects for dataset. First, we evaluated viewers' responses when answering questions using graphs of different complexity. We found that viewers produced more error when presented with graphs plotting more variables (4 variables: $\hat{\beta} = 20.372$, $t = 7.496$, $p = 7.51\text{e}{-14}$; 5 variables: $\hat{\beta} = 21.485$, $t = 7.937$, $p = 2.45\text{e}{-15}$), relative to those plotting the fewest variables (i.e., 3 variables). Viewers were also slower to provide a response when presented with graphs plotting more variables (4 variables: $\hat{\beta} = 13039.241$, $t = 9.263$, $p < 0.001$; 5 variables: $\hat{\beta} = 16264.272$,

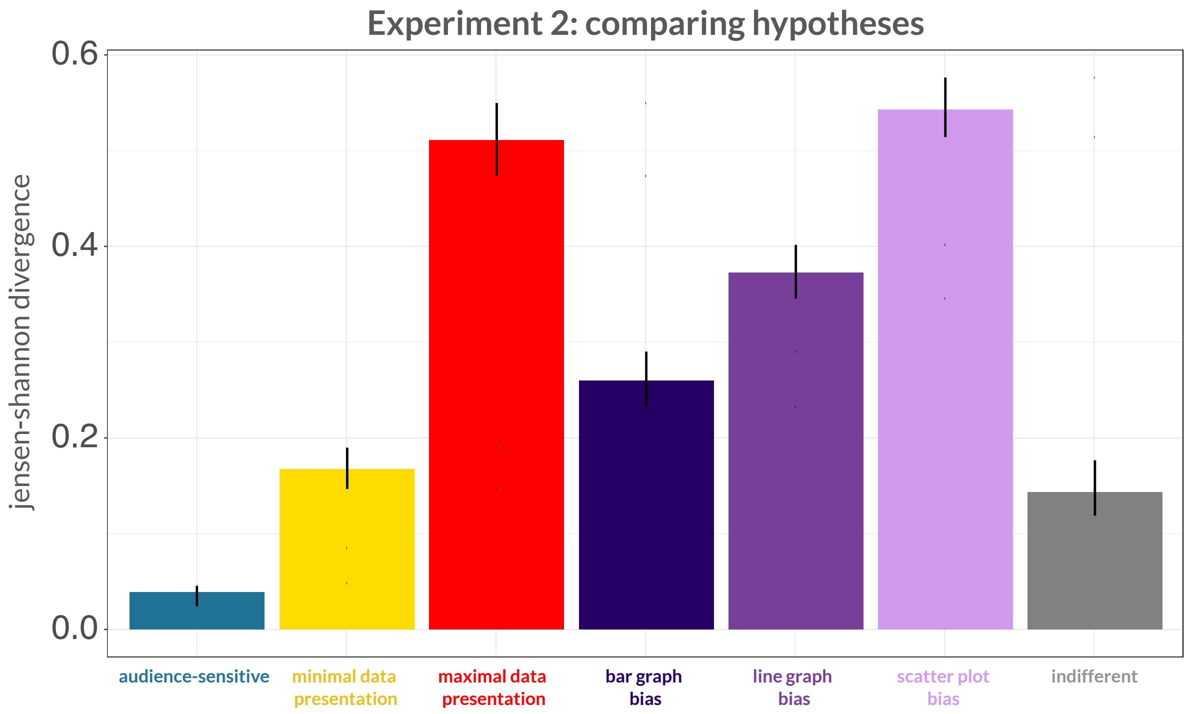


Figure 3.7. Experiment 2: Comparing hypotheses using Jensen-Shannon Divergence against participant selection behavior (identical match = 0). Error bars indicate 95% CIs.

$t = 11.6$, $p < 0.001$), compared to graphs plotting the fewest variables. These results provide evidence that simpler graphs are easier for viewers to answer questions more quickly and accurately.

Second, we evaluated how viewers responded to different graph types. We found that viewers produced more error when presented with scatter plots ($\hat{\beta} = 30.17$, $t = 11.25$, $p < 0.001$), compared to bar graphs, although they took a similar amount of time to respond to both types of graphs ($\hat{\beta} = -949.3$, $t = -0.671$, $p = 0.502$). By contrast, when responding to line graphs, viewers provided similarly accurate responses ($\hat{\beta} = 4.095$, $t = 1.51$, $p = 0.130$) and took similar amount of time to respond ($\hat{\beta} = 1106.0$, $t = 0.775$, $p = 0.438$), relative to when they were presented with bar graphs.

Lastly, we evaluated how viewers responded to the different questions. Overall, viewers produced more error when predicting trends ($\hat{\beta} = 57.98$, $t = 22.95$, $p < 0.001$), relative to comparing the means of multiple categories. Viewers also produced the least amount of error when retrieving the means of categories ($\hat{\beta} = -10.47$, $t = -4.091$, $p = 4.35\text{e}{-5}$), relative to making category mean comparisons. We found that viewers were faster to respond when asked to predict data trends ($\hat{\beta} = -4616.49$, $t = -3.301$, $p = 9.68\text{e}{-4}$) or retrieve values ($\hat{\beta} = -12024.53$, $t = -8.488$, $p < 0.001$), relative to comparing the means of categories, but were fastest to provide a response when asked to retrieve the means of categories ($\hat{\beta} = -7422.15$, $t = -4.95$, $p = 7.92\text{e}{-7}$). These results indicate that evaluating multiple categories in a graph may take more cognitive effort to produce an accurate response, regardless of its graph type or complexity. Additionally, when we added an interaction term between graph type and question type, we found that viewers produce more error when asked to predict trends using scatter plots ($\hat{\beta} = 12.96$, $t = 2.12$, $p = 3.40\text{e}{-2}$) and line graphs ($\hat{\beta} = 12.91$, $t = 2.092$, $p = 3.65\text{e}{-2}$), relative to they were presented with bar graphs. Nonetheless, viewers took a similar amount of time predicting trends across all graph types (scatter plots: $\hat{\beta} = 30.96$, $t = 0.009$, $p = 0.993$; line graphs: $\hat{\beta} = -4371.53$, $t = -1.275$, $p = 0.2025$). We also applied a likelihood ratio test to a nested model comparison, we found that variation in the graphs themselves explained additional variation in viewer responses beyond which questions they were presented with ($\chi^2(1) = 716.22$, $p < 0.001$)[2].

Leveraging these viewer responses, we then assessed the degree to which participants' selection of graphs aligned with the graphs that helped viewers answer prompted questions as accurately as possible (JSD=0.04; bootstrapped 95% CI=[0.03, 0.05]). We used the same computational procedure to evaluate how well participants selected graphs that would support faster graph comprehension and found the JSD of participants' selection behavior of 0.06 (bootstrapped 95% CI=[0.05, 0.08]). To more deeply explore these results and examine what might explain how people decide which data visualizations are most effective for specific questions, we evaluated each of our additional hypotheses predicting

[2] These results validate that our set graph stimuli in Experiment 2 were diverse enough to capture viewer response variation in our graph comprehension task.

selection bias for: (1) minimal data presentation; (2) maximal data presentation; and (3) specific graph types. For our hypothesis predicting a bias for minimal data presentation, we used a softmax decision rule that prioritized graphs plotting 3 variables ($U = 1$) and discarded graphs plotting 4 variables and graphs plotting 5 variables ($U = 0$). For our hypothesis predicting a bias for maximal data presentation, we used a softmax decision rule that prioritized graphs plotting 5 variables ($U = 1$) and discarded graphs plotting 3 variables and graphs plotting 4 variables ($U = 0$). To evaluate participants' potential biases for different graph types, we generated 3 different softmax decision rules. To model a bias towards bar graphs, a softmax decision rule prioritized bar graphs ($U = 1$) and discarded line graphs and scatter plots ($U = 0$). We used the same procedure enerating softmax decision rules prioritizing line graphs and scatter plots, respectively.

We found that our hypothesis predicting preference for minimal data presentation most aligned with participants' selections (JSD=0.17; bootstrapped 95% CI=[0.15, 0.19]), by comparison to our hypothesis predicting preference for maximal data presentation (JSD=0.51; bootstrapped 95% CI=[0.47, 0.55]) or hypothesis predicting bias for particular graph types (bar graph: JSD=0.26; bootstrapped 95% CI=[0.23, 0.29]; line graph: JSD=0.37; bootstrapped 95% CI=[0.35, 0.40]; scatter plot: JSD=0.54; bootstrapped 95% CI=[0.51, 0.58]). However, all these heuristic models fell short of aligning with our audience sensitive model for better viewer accuracy (all $p < 0.01$) and faster viewer response time (all $p < 0.01$).

Similar to Experiment 1, we conducted a split-half reliability test and also found that participants with more math experience were more variable in their graph selections (JSD=0.24) than those with less math experience (JSD=0.03). These results may explain why participants with more math experience did not necessarily choose graphs that supported more accurate (JSD=0.07; bootstrapped 95% CI=[0.04, 0.09]) and faster (JSD=0.08; bootstrapped 95% CI=[0.06, 0.12]) graph comprehension by viewers, relative to participants with less math experience (JSD accuracy = 0.04; bootstrapped 95%

CI=[0.03, 0.07], JSD response time = 0.07; bootstrapped 95% CI=[0.05, 0.10]). However, these results indicate that even participants with less math experience are sensitive to what features of a graph enable a viewer to quickly and accurately extract needed information from it in order to answer target questions.

In summary, these results indicate that while participants are biased towards more simple graphs with less variables and bar graphs, they adaptively prioritize graphs depending on the target question beyond simple heuristic strategies. Critically, these analyses provide evidence that participants across all math backgrounds prioritized graphs that support faster and more accurate viewer graph comprehension.

3.8 Discussion

Data visualization, among other tools for making sense of large volumes of data, has become increasingly important in recent decades (Holst, 2021). Here we investigated the intuitions that ordinary people have about what makes data visualizations informative for answering specific questions. Concretely, we evaluated the extent to which people may be sensitive to how communicative goals to convey different kinds of information should shift effective data visualization design.

To accomplish this, we conducted two experiments aimed at evaluating non-experts' intuitions about data visualization efficacy using two novel tasks: First, we used a forced-choice graph selection task to remove skill-based barriers associated with graph construction (e.g., manipulating data using programming languages) and to measure participants' preferences about graphs intended to communicate different kinds of information. Next, to model audience sensitivity, we developed a graph comprehension task to evaluate a separate group of naïve viewers' ability to accurately answer questions about those same graphs. In Experiment 1, we found that people prioritized bar graphs containing the minimal information needed to answer target questions, but were not necessarily sensitive to more subtle differences between how different bar graphs plotting the same data could better support fast and accurate comprehension by others. In Experiment 2, we expanded our stimuli set to vary in graph type (bar graphs, line graphs, and scatter plots) and graph complexity (3 plotted variables, 4 plotted variables, and 5 plotted variables) in order to disentangle whether participants merely attend to whether graphs have the minimal information needed to answer target questions or are sensitive to which graphs support fast and accurate viewer comprehension. We found that while participants were biased towards bar graphs and more simple graphs plotting 3 variables, they adaptively selected different graphs depending on the target question and, moreover, that their selections aligned with the graphs that supported faster and more accurate comprehension by viewers. Overall, our findings contribute quantitative evidence that even non-experts' intuitions about data visualization design are guided by goals to generate informative messages for others, despite lacking design expertise typically investigated by prior research. By leveraging viewers' downstream interpretations of the same graphs, our results additionally provide critical insights about how their design preferences have a direct downstream impact on viewers' ability to accurately extract information from graphs.

A key contribution of this work is that we establish the feasibility of systematically investigating non-experts' intuitions about data visualization design. How might their intuitions differ from those of experts? While our findings demonstrate that our participants were systematically biased to select informative graphs over uninformative ones, their selection behavior is consistent with a relatively coarse understanding of what makes a data visualization easy for someone else to understand. Whereas people are exposed to other forms of pragmatic communication like gestures and drawings and become adult-like experts in processing them from a young age (Goldin-Meadow, 2009; *Huey et al., 2022), experience with graphs typically occurs later in development and is taught in formal educational settings. Thus, a coarse understanding may develop into more fine-grained tuning as people gain more domain-specific experience with data visualizations or more broadly, mathematics. This prediction resonates with previous studies suggesting that graph reading performance can be predicted by learners' basic numerical abilities (Berg & Smith, 1994; Ludewig et al., 2020; M. J. Padilla et al., 1986), spatial reasoning about mental number lines (Booth & Siegler, 2008), and comprehension of non-symbolic and symbolic number magnitudes (Dehaene et al., 2003) and arithmetical processes (Gillan, 2009). For example, as people gain more experience with data visualizations, they may gain greater visual acuity with discerning such statistical patterns (Ali & Peebles, 2013; Ratwani & Gregory Trafton, 2008) that go beyond ingrained Gestalt Principles.

In conclusion, our paper contributes new insights about how people transform their knowledge about the world into data visualizations that others can learn from. Indeed, research of this nature investigating graphs and their presentation of statistical patterns is critical to deepening foundational understanding of communication modalities that involve symbolic reasoning, but also to the scientific community that utilizes data visualizations as a primary tool to share findings with other scientists and with the general public. Ultimately, data visualization studies guided by cognitive theories of communication may help advance the development of novel data visualization tools, as well as identify potential opportunities for graph literacy interventions in STEM education and design.

3.9 Acknowledgments

I would like to thank my co-authors for their feedback and invaluable contributions to this manuscript and earlier versions of this manuscript: Lauren A. Oey, Hannah S. Lloyd, and Judith E. Fan. And as always, I thank the members of the Cognitive Tools Lab at UC San Diego and Stanford University for helpful discussion.

Chapter 3, in full, is currently being prepared for submission for publication of the material. An earlier version of the project was published as Huey*, H., Oey*, L., Lloyd, H., & Fan, J. (2023). How do communicative goals guide which data visualizations

people think are effective? *Proceedings of the 45th Annual Meeting of the Cognitive Science Society*. Cognitive Science Society. The dissertation author was the primary investigator and author of this material.

References

Ajani, K., Lee, E., Xiong, C., Knaflic, C. N., Kemper, W., & Franconeri, S. (2021). Declutter and focus: Empirically evaluating design guidelines for effective data communication. *IEEE Transactions on Visualization and Computer Graphics*, *28*(10), 3351–3364.

Ali, N., & Peebles, D. (2013). The effect of gestalt laws of perceptual organization on the comprehension of three-variable bar and line graphs. *Human factors*, *55*(1), 183–203.

Berg, C. A., & Smith, P. (1994). Assessing students' abilities to construct and interpret line graphs: Disparities between multiple-choice and free-response instruments. *Science Education*, *78*(6), 527–554.

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

Booth, J. L., & Siegler, R. S. (2008). Numerical magnitude representations influence arithmetic learning. *Child development*, *79*(4), 1016–1031.

Börner, K., Bueckle, A., & Ginda, M. (2019). Data visualization literacy: Definitions, conceptual frameworks, exercises, and assessments. *Proceedings of the National Academy of Sciences*, *116*(6), 1857–1864.

Boy, J., Rensink, R. A., Bertini, E., & Fekete, J.-D. (2014). A principled way of assessing visualization literacy. *IEEE Transactions on Visualization and Computer Graphics*, *20*(12), 1963–1972.

Card, M. (1999). *Readings in information visualization: Using vision to think*. Morgan Kaufmann.

Clark, E. V., & Hecht, B. F. (1983). Comprehension, production, and language acquisition. *Annual Review of Psychology*, *34*(1), 325–349.

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 (General)*, *150*(3), 192–210.

Degen, J., Hawkins, R. D., Graf, C., Kreiss, E., & Goodman, N. D. (2020). When redundancy is useful: A bayesian approach to "overinformative" referring expressions. *Psychological Review*, *127*(4), 591.

Dehaene, S., Piazza, M., Pinel, P., & Cohen, L. (2003). Three parietal circuits for number processing. *Cognitive neuropsychology*, *20*(3-6), 487–506.

Fan, J. E., Hawkins, R. D., Wu, M., & Goodman, N. D. (2020). Pragmatic inference and visual abstraction enable contextual flexibility during visual communication. *Computational Brain & Behavior*, *3*(1), 86–101.

Franconeri, S. L., Padilla, L. M., Shah, P., Zacks, J. M., & Hullman, J. (2021). The science of visual data communication: What works. *Psychological Science in the Public Interest*, *22*(3), 110–161.

Franke, M., & Jäger, G. (2016). Probabilistic pragmatics, or why bayes' rule is probably important for pragmatics. *Zeitschrift für Sprachwissenschaft*, *35*(1), 3–44.

Gillan, D. J. (2009). A componential model of human interaction with graphs: Vii. a review of the mixed arithmetic-perceptual model. *Proceedings of the Human Factors and Ergonomics Society annual meeting*, *53*(12), 829–833.

Goldin-Meadow, S. (2009). How gesture promotes learning throughout childhood. *Child Development Perspectives*, *3*(2), 106–111.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829.

Grammel, L., Tory, M., & Storey, M.-A. (2010). How information visualization novices construct visualizations. *IEEE Transactions on Visualization and Computer Graphics*, *16*(6), 943–952. https://doi.org/10.1109/TVCG.2010.164

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and Semantics Vol. 3: Speech Acts* (pp. 64–75). Academic Press.

Holst, A. (2021). Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2025. *Statista, June.*

*Huey, H., *Long, B., Yang, J., George, K. R., & Fan, J. E. (2022). Developmental changes in the semantic part structure of drawn objects. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*(44).

Kao, J. T., Wu, J. Y., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, *111*(33), 12002–12007.

Kelleher, C., & Wagener, T. (2011). Ten guidelines for effective data visualization in scientific publications. *Environmental Modelling & Software*, *26*(6), 822–827.

Kosslyn, S. M. (1989). Understanding charts and graphs. *Applied Cognitive Psychology*, *3*(3), 185–225.

Lee, S., Kim, S.-H., & Kwon, B. C. (2016). Vlat: Development of a visualization literacy assessment test. *IEEE Transactions on Visualization and Computer Graphics*, *23*(1), 551–560.

Lee, S., Kwon, B. C., Yang, J., Lee, B. C., & Kim, S.-H. (2019). The correlation between users' cognitive characteristics and visualization literacy. *Applied Sciences*, *9*(3), 488.

Leonard, J. G., & Patterson, T. F. (2004). Simple computer graphing assignment becomes a lesson in critical thinking. *NACTA Journal*, 17–21.

Ludewig, U., Lambert, K., Dackermann, T., Scheiter, K., & Möller, K. (2020). Influences of basic numerical abilities on graph reading performance. *Psychological Research*, *84*, 1198–1210.

Mansoor, H., & Harrison, L. (2018). Data visualization literacy and visualization biases: Cases for merging parallel threads. *Cognitive Biases in Visualizations*, 87–96.

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), 1–25.

Padilla, M. J., et al. (1986). An examination of the line graphing ability of students in grades seven through twelve. *School Science and Mathematics*, *86*(1), 20–26.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*(4), 329–347.

Pinker, S. (1990). A theory of graph comprehension. *Artificial Intelligence and the Future of Testing*, *73*, 126.

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

Rensink, R. A., & Baldridge, G. (2010). The perception of correlation in scatterplots. *Computer Graphics Forum*, *29*(3), 1203–1210.

Saket, B., Moritz, D., Lin, H., Dibia, V., Demiralp, C., & Heer, J. (2018). Beyond heuristics: Learning visualization design. *arXiv preprint arXiv:1807.06641*.

Shah, P., & Hoeffner, J. (2002). Review of graph comprehension research: Implications for instruction. *Educational Psychology Review*, *14*, 47–69.

Tufte, E. R. (1983). The visual display of quantitative information.

Tversky, B. (2001). Spatial schemas in depictions. *Spatial schemas and abstract thought*, *79*, 111.

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (Fourth) [ISBN 0-387-95457-0]. Springer. https://www.stats.ox.ac.uk/pub/MASS4/

Wang, Y., Han, F., Zhu, L., Deussen, O., & Chen, B. (2017). Line graph or scatter plot? automatic selection of methods for visualizing trends in time series. *IEEE transactions on visualization and computer graphics*, *24*(2), 1141–1154.

Chapter 4

Discussion

How do task goals shift the kinds of visualization strategies that people use to communicate their knowledge? Throughout my dissertation, I explore the idea that visual communication is not a strictly reconstructive process, but is instead a cognitive process of balancing what considered perceptually salient and semantically relevant to represent that is guided by pragmatic inferences about what a viewer may need to gain from the visualization. To do this, I developed novel visual production tasks and a series of viewer interpretation tasks to: (1) explore what visualization strategies people use to convey different ideas to viewers; (2) evaluate how task goals may shift what information people choose to prioritize in their visualizations; and (3) measure how these how these representational choices directly impact downstream interpretation by viewers, depending on their own task goals.

In Chapter 1, I leveraged a drawing task for studying visual communication of visual object concepts at different levels of abstraction. I also introduced a novel semantic annotation task in which a second group of participants provided fine-grain annotations of what each stroke in our generated drawing dataset of $>$ 12K drawings of 32 object concepts represented. These tools enabled me to conduct a detailed investigation of the semantic part-level information that people prioritized in their drawings depending on their task goals (i.e., to convey exemplar-level or category-level information) and immediate sensory inputs (i.e., being cued by a category label or photograph), as well as how these different prioritizations of visual information impacted each drawing's recognizability to naive viewers. We found that drawings meant to convey exemplars of visual object concepts (e.g., "Garfield") are easier to identify but less categorizable than drawings meant to convey category-level concepts (e.g., "cat"), suggesting that people prioritize different diagnostic information in their drawings when drawing visual object concepts at varying levels of abstraction. Moreover, drawings that were cued by photograph are less categorizable than drawings cued by category label, although this gap is reduced when the photograph is of a more typical exemplar. These data provide a nuanced understanding of how drawings encode meaning at different levels of semantic abstraction, suggesting a dissociation between how drawings communicate more general vs. more specific meanings. Taken together, this work demonstrates both the flexibility of people's visual communication strategies and, moreover, that their behaviors are systematically guided by communicative goals and immediate sensory inputs to jointly determine the kind of visual information that people include in their drawings.

Chapter 2 builds on this investigation of flexible visual production behaviors by systematically evaluating how task goals to produce visual explanations of how object function differ from depictions of what objects look like. Using the same drawing and annotation paradigms developed in Chapter 2, I found that people emphasized functionally critical parts of these objects when producing visual explanations, using more strokes to draw these parts and making them appear larger than when they only aimed to produce a visually accurate drawing of the object. They also selectively included abstract symbols in their visual explanations, including arrows and motion lines, suggesting that they believe that providing an explanation means going beyond drawing physical components of the same object. While these explanatory drawings more effectively communicated which action was needed to interact with the object than depictive drawings, this enhancement was accompanied by a loss in diagnostic information about the object's visual appearance. Taken together, these findings suggest how people produce visual explanations is systematic, prioritizing information about function (i.e., how parts move and interact) over information about structure (i.e., what parts look like and where they are). In sum, Chapters 1 and 2 provide evidence that, although people use flexible and adaptive visualization strategies to convey diverse ideas spanning different levels of abstraction, their strategies are systematic depending on their task goals. Moreover, how these task goals constrains how people generate drawings and visual explanations has direct downstream consequences on viewers' interpretations.

While Chapters 1 and 2 leverage drawing—one of the most versatile and accessible visualization technologies (Eitz et al., 2012; Fan et al., 2018; Sangkloy et al., 2016)—to investigate visual communication behaviors, Chapter 3 assesses the extent to which these flexible visualization strategies extend to more domain-specific experience such as with data visualizations or more broadly, mathematics. Unlike drawings, data visualizations are a much more modern visualization technique and often requires computational tools and expertise to generate them. These barriers, as well as the mathematical knowledge to understand data visualizations, make it unclear whether people have strong intuitions about what makes data visualizations more or less informative for graph comprehension by viewers. The studies in Chapter 3 tackle this challenge by developing a novel forced-choice graph selection task that allows researchers to test data visualization intuitions of everyday people in a communicative setting by removing skill-based barriers, like manipulating data using programming languages. Furthermore, by developed a graph comprehension task, I developed estimates for a baseline for how well the graphs in the studies actually communicate information, by querying a separate group of participants to extract information from graphs, with varying degrees of accuracy. These studies suggest that people are sensitive to selecting graphs that can effectively convey critical quantitative patterns to viewers. With data visualizations being one of the most computationally complex of visualizations, these insights help develop more unified theories of visual communication aimed at uncovering the cognitive mechanisms underlying how we encode and convey our knowledge of the world to others through visual form.

In conclusion, my dissertation synthesizes large-scale crowdsourcing techniques and novel visual production tasks to evaluate how people communicate their knowledge through visual form. Building on growing work investigating depiction as a communicative act adaptive to different contexts (Fan et al., 2020; Galantucci, 2005; Garrod et al., 2007), the studies in this dissertation provide evidence that the act of visual communication relies on rich interactions between perception and pragmatics that guide how people flexibly but systematically encode through visualizations like drawings, diagrams, and data visualizations. Moreover, people selectively prioritize different kinds of visual information depending on their task goals. These insights resonate with communicative research across linguistics (Clark & Wilkes-Gibbs, 1986; Hawkins et al., 2017) and gesture (Goldin-Meadow, 2005; Goldin-Meadow et al., 1996; Sandler et al., 2005). Building on these prior studies, the results of my dissertation help contribute towards a more unified theory of visual communication that is analogous to linguistic theories of communication. Moreover, by evaluating how well viewers can interpret these visualizations produced under different task contexts, my work bridges how people's representational choices directly impact how viewers interpret how portrayed visual information relate to target objects, scenes, and events.

Additionally, a major contribution of my dissertation is publicly offering all datasets of this research. By systematically measuring the visual information in these visualizations and as well as systematically manipulating the task contexts under which they were produced, these benchmark datasets of human generative and interpretational behaviors can help inform modern automatic generative text-to-visual systems (e.g., FireFly, Midjourney, Dall-E, ChatGPT-4). Overall, these insights open new opportunities to explore the nature of how people map the correspondence between their internal and external representations of visual object concepts. Although my dissertation only scratches the surface of our understanding of how people communicate their knowledge through visual form—whether to catalyze some of human history's most impressive scientific discoveries (e.g., through visualizations like the periodic table, coordinate system, Vitruvian man) or now to collaborate with artificially intelligent agents to automatically generate novel visualizations—perhaps that's how it all began. With just a few markings.

References

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39.

Eitz, M., Hays, J., & Alexa, M. (2012). How do humans sketch objects? *ACM Transactions on graphics (TOG)*, *31*(4), 1–10.

Fan, J. E., Hawkins, R. D., Wu, M., & Goodman, N. D. (2020). Pragmatic inference and visual abstraction enable contextual flexibility during visual communication. *Computational Brain & Behavior*, *3*(1), 86–101.

Fan, J. E., Yamins, D. L., & Turk-Browne, N. B. (2018). Common object representations for visual production and recognition. *Cognitive science*, *42*(8), 2670–2698.

Galantucci, B. (2005). An experimental study of the emergence of human communication systems. *Cognitive science*, *29*(5), 737–767.

Garrod, S., Fay, N., Lee, J., Oberlander, J., & MacLeod, T. (2007). Foundations of representation: Where might graphical symbol systems come from? *Cognitive science*, *31*(6), 961–987.

Goldin-Meadow, S. (2005). *The resilience of language: What gesture creation in deaf children can tell us about how all children learn language*. Psychology Press.

Goldin-Meadow, S., McNeill, D., & Singleton, J. (1996). Silence is liberating: Removing the handcuffs on grammatical expression in the manual modality. *Psychological Review*, *103*(1), 34.

Hawkins, R. X., Frank, M., & Goodman, N. D. (2017). Convention-formation in iterated reference games. *CogSci*.

Sandler, W., Meir, I., Padden, C., & Aronoff, M. (2005). The emergence of grammar: Systematic structure in a new language. *Proceedings of the National Academy of Sciences*, *102*(7), 2661–2665.

Sangkloy, P., Burnell, N., Ham, C., & Hays, J. (2016). The sketchy database: Learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)*, *35*(4), 1–12.