UNIVERSITY OF CALIFORNIA SAN DIEGO

On the importance of context: How auxiliary information from within- and across-modalities guides, facilitates, and perturbs visual processing

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

in

Experimental Psychology

by

Jamal R. Williams

Committee in charge:

Professor Timothy Brady, Chair
Professor Viola Störmer, Co-Chair
Professor Sarah Creel
Professor John Serences

2024

The Dissertation of Jamal R. Williams 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 amazing partner Lucia Trapote. Thank you for your compassion and unwavering support through some of the most challenging moments in my life.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

None of the work contained in this dissertation would have been possible without my partner Lucia Trapote. I am deeply grateful for your kindness, your support, your love, for reviewing all of my papers and presentations, and for being the first pilot subject on many of much of my work. Thank you.

I owe an immense debt of gratitude to my exceptional advisors, Tim Brady and Viola Störmer. Your invaluable mentorship has been instrumental in shaping me into the scientist I am today. Through your guidance, I have navigated countless challenges and celebrated many successes. I owe a great deal of my accomplishments to your unwavering support. Your patience and understanding have made all of this work possible, and I am forever indebted to you for going above and beyond what I could have ever expected of a mentor. I would also like to thank the rest of my wonderful committee: John Serences and Sarah Creel for their thoughts and advice which have made this and other work better.

I am very grateful for having such a supportive lab throughout this process. I'd like to thank my "grad siblings" Hayden Schill and Isabella Distefano who started the program with me and helped elevate my work. I'd like to thank Chaipat Chunharas for being my only office mate in five years, and for demonstrating what it means to make an engaging, meaningful, and accessible presentation. I am also lucky to have shared thoughts and collaborated with Maria Robinson, Yong Hoon Chung, Johnathan Keefe, Angus Chapman, Michael Allen, Mark Schurgin, and my wonderful research assistants: Avery Quynh, Delaney Pickell, and Lulu Ricketts. My success has been dependent on too many people thank and, to everyone who wasn't named but should have been, I am truly sorry. Lastly, I'd like to thank whoever read and approved my NSF GRFP application. I am extremely lucky to have won that award as this funding made much of this work possible and which made it possible to spend several months at Dartmouth to continue and expand my work with the Störmer lab.

Chapter 1, in full, is a reprint of the material as it appears in the Journal of Experimental Psychology: Human Perception and Performance, 2022, Williams, Jamal R.; Brady, Timothy F.; Störmer, Viola S. The dissertation author was the primary investigator and author of this paper.

Chapter 2, in full, is a reprint of the material as it appears in Psychological Science, 2024, Williams, Jamal R.; Störmer, Viola S. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is a reprint of the material as it appears in Psychological Science, 2022, Williams, Jamal R.; Markov, Yuri A.; Tiurina, Natalia A.; Störmer, Viola S. The dissertation author was the primary investigator and author of this paper.

VITA

2018 Bachelor of Science in Cognitive and Behavioral Neuroscience, University of California San Diego

2020 Master of Arts in Experimental Psychology, University of California San Diego

2024 Doctor of Philosophy in Experimental Psychology, University of California San Diego

PUBLICATIONS

Williams, J. R., & Störmer, V. S. (accepted, 2024). Cutting through the noise: Auditory scenes and their effects on visual object processing Psychological Science

Brady, T. F., Robinson, M. M., & Williams, J. R. (2024). Noisy and hierarchical visual memory across timescales. Nature Reviews Psychology, 3(3), 147–163.

Brady, T. F., Robinson, M. M., Williams, J. R., & Wixted, J. T. (2023). Measuring memory is harder than you think: How to avoid problematic measurement practices in memory research. Psychonomic Bulletin & Review, 30(2), 421–449.

Williams, J. R., Robinson, M. M., Schurgin, M. W., Wixted, J. T., & Brady, T. F. (2022). You cannot "count" how many items people remember in visual working memory: The importance of signal detection–based measures for understanding change detection performance. Journal of Experimental Psychology: Human Perception and Performance, 48(12), 1390–1409.

Williams, J. R., Markov, Y. A., Tiurina, N. A., & Störmer, V. S. (2022). What You See Is What You Hear: Sounds Alter the Contents of Visual Perception. Psychological Science, 33(12), 2109–2122.

Williams, J. R., Robinson, M. M., & Brady, T. F. (2022). There Is no Theory-Free Measure of "Swaps" in Visual Working Memory Experiments. Computational Brain & Behavior.

Williams, J. R., Brady, T. F., & Störmer, V. S. (2022). Guidance of attention by working memory is a matter of representational fidelity. Journal of Experimental Psychology: Human Perception and Performance, 48(3), 202–231.

Brady, T.F., Störmer, V.S., Shafer-Skelton, A., Williams, J.R., Chapman, A.F., & M. Schill, H. (2019). Scaling up visual attention and visual working memory to the real world. Psychology of Learning and Motivation - Advances in Research and Theory, 70, 29–69.

ABSTRACT OF THE DISSERTATION

On the importance of context: How auxiliary information from within- and across-modalities guides, facilitates, and perturbs visual processing

by

Jamal R. Williams

Doctor of Philosophy in Experimental Psychology

University of California San Diego, 2024

Professor Timothy Brady, Chair
Professor Viola Störmer, Co-Chair

Despite the sense of a rich and complete visual world, we are only capable of processing a small fraction of the sensory information available to us. It is no wonder then, that achieving this level of perception involves a series of complex operations from the moment visual information first hits the retina to its eventual interpretation in higher regions of the brain. In this dissertation I explore how visual and auditory information might be leveraged by the visual system to predict impending stimulation and alleviate this burden. In chapter 1, I investigate how the contents of visual working memory captures attention. In a series of studies, I demonstrate that this automatic co-opting of attention is driven by the fidelity of the internal representation, that this capture occurs even when working memory is taxed, and that the quality of any given internal representation is driven by a natural and stochastic accumulation of noise in the representation itself. In the following chapters I explore how real-world auditory objects and scenes facilitate (chapter 2) and alter visual perception (chapter 3). Within these chapters I employ a visual discrimination task where I control the unraveling of visual information to effectively “slow down” visual processing. In chapter 2, I demonstrate that people are able to discern meaningful information more rapidly when the auditory objects and scenes are semantically related to their visual target compared to when they are not. In chapter 3, I demonstrate that this seemingly facilitatory process is not cost free. When the visual targets are less discrete, and more ambiguous, auditory information not only accelerates visual processing but, in so doing, alters the perceptual representation of visual objects, shifting them towards expected features associated with the accompanying sounds. Importantly, this perturbation is not driven by decisional processes, nor is it driven by volitional attentional selection. In summary, this dissertation advances our understanding of visual perception by elucidating the interplay between contextual information across modalities and visual perception. By revealing how auxiliary information influences visual attention and perception, it provides insights into the mechanisms underlying the complex process by which we make sense of the visual world.

INTRODUCTION

The apparent ease with which we see belies the intricate processes necessary for visual perception to unfold. Given the nature of the (comparatively impoverished and two-dimensional) information that hits our retina, it is truly remarkable that we are capable of perceiving a rich and three-dimensional world. Achieving this, however, is metabolically and computationally intensive. From the moment visual input hits the retina until its final interpretation in the brain, visual information is heavily processed by numerous brain structures and millions of neurons (Bruce et al., 2014; Hubel & Weisel, 1979; Palmer, 1999; Stryer, 1996). Even still, once the raw input is processed, visual perception demands further processing of this input (Kersten et al., 2004; Wörgötter et al., 2004). It is no wonder then, that for visual perception to occur, the brain dedicates a substantial amount of real-estate and computational power for processing this already preprocessed information (Hubel & Weisel, 1979; Bullier, 2001; Marr, 2010). In this dissertation I will argue that visual perception inherently incorporates prior experiences and concurrently available context to predict impending visual input in an effort to alleviate the immense burden of visual processing.

The field of vision science has made substantial progress in the last several decades but attempts to understand vision and visual perception have been attempted for millennia. Since at least the 8th century B.C., philosophers have wondered why we cannot see color at night or why the moon appears large on the horizon and small at its apex (Crone, 2012; Sedley, 2018). Complex and thorough theories of visual perception have been posited since at least the 5th century. Empedocles put forth the theory that light is emitted from the eyes like a lantern and produces visual perception as it illuminates anything it touches (O'Brien, 1970). And while this theory would not be completely dismissed for over a thousand years (*see* Al-Khalili, 2015 *and*

Burton, 1945), during the 4th century BC, Epicurus and others rejected this “emission” theory and proposed one of “intromission” whereby objects emit “atoms” in the form of “essential waves” that make physical contact with the eye (Crone, 2012; Hahm, 1978; Løkke, 2008). As our understanding of the physical world improved, and we moved beyond “emission” theories (al-Haytham, 1021; Newton, 1704), scientists questioned how our rich perception could be derived from impoverished and indeterminate visual input (Craik, 1967; Koffka, 1935; Helmholtz, 1867).

Recognizing the clear disparity between input and perception, many theories how we perceive the visual world arose. One influential example, *Gestalt* theory, posited several unassailable principles of perception, such as continuity, common fate, emergent properties, and more (*see,* Koffka, 1922). As seen in Figure 1A, even though each dot is independent, an emergent property of their configuration gives the sense of curvature, orientation, and of belonging to a larger whole which is not shared by the components (Palmer, 1999). While this school of thought made important and lasting contributions to the field, it has generally been surpassed by newer theories and models of visual perception (*for comprehensive reviews of this school and its contributions, see* Wagemans et al., 2012a; Wagemans et al., 2012b). Others, like the theory of *unconscious inference*, have survived in one form or another and continue to inform theories of visual perception. According to this theory, the inherent ambiguity in visual information is overcome through implicit predictions on the source of visual stimulation and these unconscious inferences must be integrated somewhere in the visual processing system to generate perception (Helmholtz, 1867; Kersten et al., 2004; Knill & Pouget, 2004; Stokes et al., 2012).

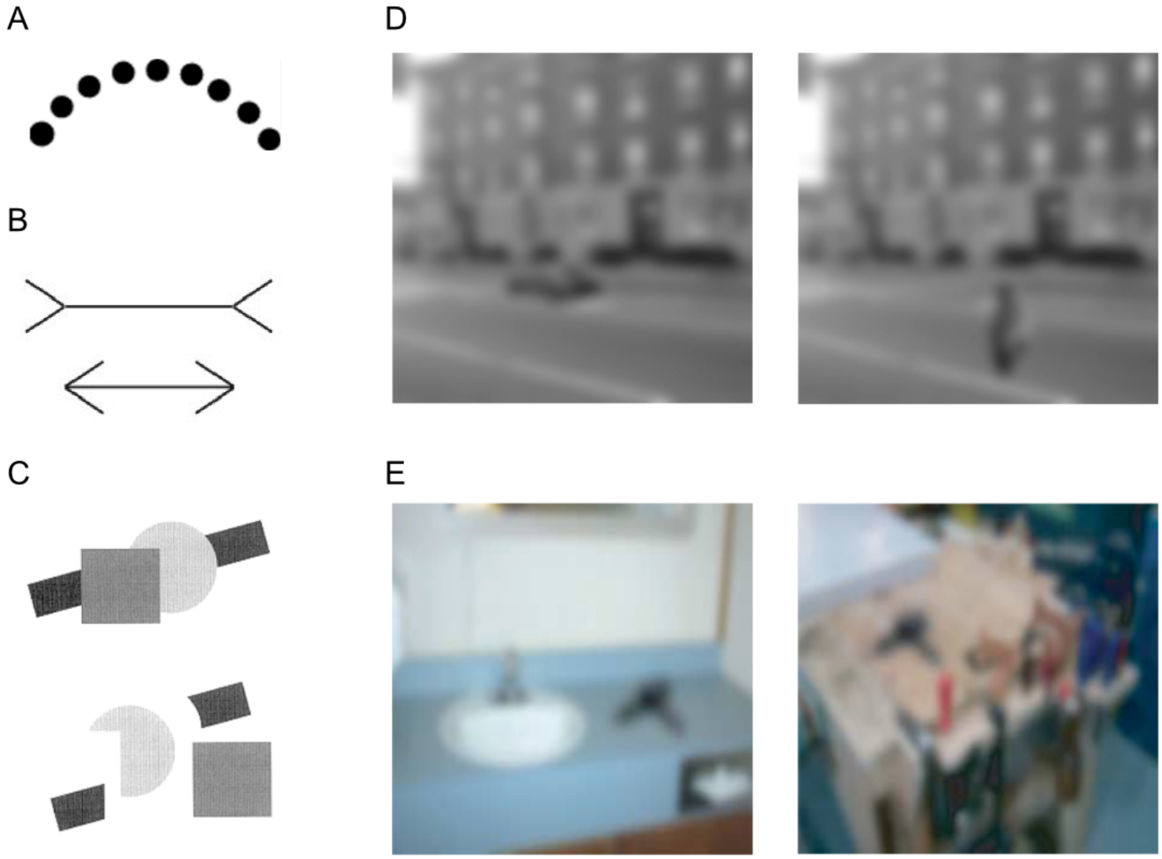


Figure 0.1 Unconscious inference at work

Several examples of classic and recent findings which elucidate the unconscious inferences that are applied to ambiguous visual information. (A) An example of the Gestalt principle of emergent properties. (B) The classic Müller-Lyer illusion, both lines are identical, yet it is common to perceive the top line as longer than the bottom. (C) A simple example of amodal completion, the top arrangement implies full shapes that are simply occluded, while the actual visual stimulation intuitively feels much less likely. (D-E) Examples of how a scene or an object's position and orientation can give rise to different inferences on object identity, for example (from top left, clockwise) a car, a person, a drill, a hairdryer.

While the theory of unconscious inference was not entirely embraced at the time, today, it is well accepted[1] that visual perception is not strictly stimulus driven (Craik, 1967; Kersten et al.,

[1] The phrase "unconscious inference" has fallen out of favor, but the principles are commonplace and are often reframed simply as inference (Bruner & Potter, 1964), perceptual inference (e.g., Parise, 2016; Rohe & Noppeney, 2015) or Bayesian inference (e.g., Kersten et al. 2004; Knill & Pouget, 2004; Pouget et al., 2013).

2004; Kinchla & Wolfe, 1979; Koffka, 1922; Koffka, 1935; Palmer, 1999; Torralba, 2003). As a demonstration, in Figure 1B, the top line often appears longer than the bottom line, even though the target visual information (i.e., line length) is identical across samples (Lewis, 1908). One compelling reason for why this classic Müller-Lyer illusion works is that it is driven by context and priors: when edges interact like this in the world (*context*), the lines that we perceive as longer, are often genuinely longer than when we encounter similar interactions to the sample below (*prior*; see Howe and Purves, 2005). Similarly, in Figure 1C (top) it is not hard to intuit that the image is likely made up of whole objects that are simply layered on top of one another (Gerbino & Salmaso, 1987; Sekuler & Palmer, 1992). However, if we decompose the actual visual information we are shown (presented below, Figure 1C, bottom), the configuration feels odd perhaps because it is far less likely to be the true source of the visual information in the real world (Michotte et al., 1991). This act of "filling in the gaps" and completing the shapes (i.e., amodal completion) typically presumes that we assume the simplest organization possible (*but see,* Singh, 2004; Moravec & Beck, 1986).

These sorts of inferences over visual ambiguity occur at all levels of the visual hierarchy. As in the Muller-Lyer illusion, the visual system takes advantage of the statistical regularities in the world for higher level stimuli like objects, where people often interpret an object differently based on context: people often see an object as a car or a hair dryer (Figure 1D & 1E, left) and will interpret the exact same object as a person or a drill when its position in the scene, or the scene itself is changed (Figure 1D & 1E, right; Oliva & Torralba, 2007; Bar, 2004). Even when information is no longer available for continued sensory sampling, the contents of visual working memory act as a sort of context, providing information on relevant features or objects in the environment[2]. When a single visual feature or simple object is actively maintained in visual working memory, attention is rapidly and automatically deployed to any location in the environment that matches those features (Olivers et al., 2006; Soto et al., 2005). In chapter 1, I explore how this task-irrelevant, yet actively maintained, visual information can influence visual processing and attentional selection. In particular we explore whether the contents of working memory need to have an elevated status to interact with attention and whether this sort of attention capture occurs when multiple items are actively maintained.

In the subsequent chapters I explore how, similarly, task-irrelevant, but concurrently available, information might facilitate or bias active visual processing. Critically, in these chapters, I move beyond strictly visual-to-visual interactions and investigate how sounds might affect visual processing, cross-modally. It stands to reason that if the visual system incorporates secondary information to resolve ambiguity that information from distinct modalities might be incorporated as well (Parise, 2016; Rohe & Noppeney, 2015). However, the strength and directionality of these interactions is unclear. Typically, vision is thought to dominate other senses, for example, we often believe that a ventriloquist's dummy is speaking even though the spatio-auditory information suggests otherwise: visually the extravagant movements of the dummy imply that it is the source. Similarly, when visual and auditory information are equivalently ambiguous, vision is seen to dominate the auditory experience as in the classic McGurk effect (McGurk & MacDonald, 1976). In contrast, this intuition of visual dominance, and the findings that support it, might have unduly put audition in the more ambiguous situation. If so, a rational system might be expected to favor sensory information that is more precise or

[2] While working memory and long-term memory likely operate over the same representations (*see* Brady et al., 2024), working memory deserves special mention here since it is implicitly presumed that priors—the generalized sum of our experiences—exist within long-term memory (e.g., Hemmer & Steyvers, 2009).

more reliable (e.g., Körding et al., 2007; Burr et al., 2009). If visual information were more ambiguous, auditory information dominates vision (Shams et al., 2000; Alais & Burr, 2004). Therefore, to explore how audition might affect vision, we provide clear and unambiguous auditory information while presenting either (1) noisy and unambiguous information (chapter 2) or (2) noisy and ambiguous information (chapter 3).

Attentional guidance by remembered visual information

Chapter 1 investigates whether visual attention is captured by colors in the environment that match those being actively maintained in visual working memory. For example, when looking for a friend in a crowd, we may carefully hold their distinctive visual features in mind—like the red shirt they are wearing—as we try to find them. And, while our attention will often be guided toward matching features in the environment, what if our memory is inaccurate and your friend's shirt is actually more orange than red? Surprisingly, while attentional guidance has been studied extensively, the relationship between the fidelity of a memory representation and how effectively that item can guide attention is not well understood. Instead, most work has focused on the number of items that can guide attention and whether such items require a special status within working memory (e.g., by being attended within the "focus of attention").

Importantly, when only a single feature is maintained in working memory, attention is reliably and automatically guided toward matching features in the environment (Olivers et al., 2006; Soto et al., 2005, 2008; Soto & Humphreys, 2007, 2008). However, it is less clear whether multiple working memory items can guide attention in a similarly incidental way. This inability to guide when working memory is loaded beyond a single item is typically explained as the inability of multiple items to maintain a special status within working memory (e.g., Olivers et al., 2006; van Moorselaar et al., 2014). While some work has shown that multiple items can guide attention, a significant literature suggests that guidance by multiple memory representations is at best more fragile and less reliable than guidance by a single item, as guidance effects are often not found when participants remember more than a single item (Frātescu et al., 2019; van Moorselaar et al., 2014; see Ort & Olivers, 2020, for a review). In this chapter, we propose that differing results in the literature are accounted for primarily by variation in memory strength (i.e., how well information in working memory represents the initially encoded item on average).

Overall, across several experiments, we find evidence to support an account where items vary naturally in their representational fidelity, and that any and all memory items can guide attention insofar as they are well represented—even when they do not possess a special template status. Our results thereby unify the seemingly irreconcilable findings that one or many working memory items can guide attention: When working memory resources are stretched among multiple active representations, often only a single item is represented well enough to guide; however, in other cases, the fidelity of multiple items may be precise enough to produce guidance from both items.

Disambiguation by auditory objects and scenes

In chapter 2, we pivot from unimodal work and explore whether concurrently available, and task-irrelevant sounds affect active visual processing. In the real world, sounds are inexorably linked to the objects that generate them. Cats cannot bark and toads do not roar. In a world where an object's visual features such as colors and orientations are inconsistent across viewpoints, lighting conditions, and time—where visual objects are often occluded and where many objects share similar visual features despite being fundamentally distinct—our perceptual system is required to constantly make inferences about the world (Alais & Burr, 2004; Bar, 2004; Körding et al., 2007; Oliva & Torralba, 2007). Context can help us to disambiguate indeterminate information: for example, the same shape projected on our retina might be interpreted as a hair dryer when viewed in a bathroom scene or as a drill when viewed on a workbench (Bar, 2004; Biederman et al., 1982). Similarly, visual scenes can facilitate the recognition of these objects quite dramatically (Davenport & Potter, 2004; Draschkow & Võ, 2017; Palmer, 1975). In the real world, however, sensory processing of visual scenes and visual objects is highly correlated, and if information from this modality is unclear, a scene is unable to provide additional, independent information (e.g., at dusk, all visual inputs are equally obscured). In this case, nonvisual information, such as sounds, might provide unambiguous and independent information about visual inputs, and potentially influence object recognition (Plass et al., 2017). How, then, might naturalistic sounds influence the recognition of visual objects? And can both auditory objects and auditory scenes affect visual object processing?

In our study, participants viewed noisy visual objects while listening to naturalistic sounds. When sounds were related to the visual target, they facilitated the ability to extract relevant visual information, thereby accelerating object recognition. This was true for specific object sounds (a dog's bark) but also occurred for ambient auditory scene sounds (an airport terminal), indicating wide-ranging effects of audition on vision. Crucially, sounds aided categorical visual recognition (a dog from a bird) but also aided fine-grained visual discrimination (e.g., a malamute from a husky). Overall, our results demonstrate that sounds enhance vision across various levels of processing and stress the importance of cross-modal influences on perception.

Perceptual alteration by auditory objects

Chapter 3 explores the mechanisms that drive the cross-modal, facilitatory effect observed in chapter 2. While chapter 2 showed that cross-modal integration can facilitate visual processing, in chapter 3, we explore the mechanisms that drive the facilitatory effect and explore whether this effect alters our phenomenology of visual objects or simply increases the speed by which visual objects are processed (without any change to the resultant perception). As a thought experiment, imagine you catch a glimpse of something rapidly flying by your window. Because the visual information was ambiguous and fleeting, it could be any number of things. In this scenario, auditory information could be incredibly useful for resolving this uncertainty: A buzzing would suggest it was a drone, whereas a caw suggests it was a crow. Does the sound of a drone change our visual experience and make this dubious object appear more drone-like than if we hadn't heard the sound? Or do sounds simply improve perceptual processing of related visual objects by speeding responses or improving accuracy.

In chapter 3, we show that object representations are shifted away from what was presented and towards the visual features that align with the sound. These findings demonstrate that what we hear has profound impacts on how we perceive the visual world, and that the facilitation of related visual information might be driven by a form of expectation related assumptions made by the visual system. Because the influence of sound on vision seems particularly effective when visual information is noisy or dubious—where sounds provide independent and unequivocal clues about the visual environment (Alais & Burr, 2004; Heron et al., 2004; Rohe & Noppeney, 2015; Watanabe & Shimojo, 2001)—we used ambiguous visual stimuli paired with clear and distinct sounds. Specifically, we created a set of ambiguous visual stimuli by morphing together the features of two visual objects (objects A and B, e.g., a hammer and a seal) and presented these stimuli with naturalistic sounds that were congruent with one of these progenitor objects. Visual objects and sounds were presented simultaneously, and participants looked for a target object in visual noise, after which they precisely reported that object using a continuous report method.

Our results suggest that naturalistic auditory information alters the representations of objects we see. Specifically, we found that visual features of object representations are shifted toward features that are congruent with a concurrent auditory stimulus: The same ambiguous object (e.g., a 50% seal and 50% hammer morph) was perceived as more hammer-like when paired with a hammer sound and more seal-like when paired with the sound of seal barking`. In a series of control experiments, we also tested at what processing stage these audiovisual effects arose and found evidence consistent with the hypothesis that the effects of sounds on visual object recognition have an early, perceptual locus.

In sum, these chapters represent a significant advancement in our understanding of how auxiliary information affects visual attention and perception. I show that actively maintained visual representations can incidentally capture and guide our visual attention towards matching contents are present in the environment—even when multiple representations are concurrently active. I also show that specific and ambiguous auditory information in the form of object and scene sounds influence visual perception by accelerating how quickly relevant visual features are extracted from noisy visual input. Lastly, I demonstrate that this facilitation is not cost free; instead, the complete perceptual representation may be slightly shifted away from veridical and towards the expected visual features of a noisy and ambiguous visual object. Across three chapters I demonstrate that visual processing can leverage the context from multiple senses to guide, facilitate, and even perturb perception and attention; likely in an effort to lessen the intense burden required to process the massive amount of visual information at any given moment.

REFERENCES

Al-Khalili, J. In retrospect: Book of Optics. *Nature* **518**, 164–165 (2015). https://doi.org/10.1038/518164a

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. Current Biology, 14(3), 257–262. https://doi.org/10.1016/j.cub.2004.01.029

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience*, *5*(8), 617–629. https://doi.org/10.1038/nrn1476

Brady, T. F., Robinson, M. M., & Williams, J. R. (2024). Noisy and hierarchical visual memory across timescales. *Nature Reviews Psychology*, *3*(3), 147–163. https://doi.org/10.1038/s44159-024-00276-2

Burr, D., Banks, M. S., & Morrone, M. C. (2009). Auditory dominance over vision in the perception of interval duration. *Experimental Brain Research*, *198*(1), 49–57. https://doi.org/10.1007/s00221-009-1933-z

Burton, H. E. (1945). The Optics of Euclid (English translation from Latin). *Journal of Optical Society of America*, *35*(5).

Bullier, J. (2001). Integrated model of visual processing. *Brain Research Reviews*, *36*(2–3), 96–107. https://doi.org/10.1016/S0165-0173(01)00085-6

Bruce, V., Georgeson, M. A., Green, P. R., & Georgeson, M. A. (2014). *Visual Perception: Physiology, Psychology and Ecology* (4th ed.). Psychology Press. https://doi.org/10.4324/9780203427248

Bruner, J. S., & Potter, M. C. (1964). Inference in Visual Recognition. *Science*, *144*(3617), 424–425.

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. Cognitive psychology, 14(2), 143-177.

Craik, K. J. W. (1967). *The nature of explanation* (Vol. 445). CUP Archive.

Crone, R. A. (2012). *A history of color: the evolution of theories of light and color*. Springer Science & Business Media.

Davenport, J. L., & Potter, M. C. (2004). Scene consistency in object and background perception. Psychological Science, 15(8), 559–564. https://doi.org/10.1111/j.0956-976.2004.00719.x

Draschkow, D., & Võ, M. L.-H. (2017). Scene grammar shapes the way we interact with objects, strengthens memories, and speeds search. Scientific Reports, 7(1), Article 16471. https://doi.org/10.1038/s41598-017-16739-x

Frătescu, M., Van Moorselaar, D., & Mathôt, S. (2019). Can you have multiple attentional templates? Large-scale replications of Van Moorselaar, Theeuwes, and Olivers (2014) and Hollingworth and Beck (2016). Attention, Perception, & Psychophysics, 81(8), 2700–2709. https://doi .org/10.3758/s13414-019-01791-8

Gerbino, W., & Salmaso, D. (1987). The effect of amodal completion on visual matching. *Acta Psychologica*, *65*(1), 25–46. https://doi.org/10.1016/0001-6918(87)90045-X

Giard, M.-H., & Peronnet, F. (1999). Aditory-Visual Integration during Multimodal object recognition in humans: A behavioral and electrophysiological study. *Journal of Cognitive Neuroscience*, *11*(5), 473–490.

Hahm, D. E. (1978). Early Hellenistic theories of vision and the perception of color. *Studies in Perception*, 60-95.

Hemmer, P., & Steyvers, M. (2009). Integrating episodic memories and prior knowledge at multiple levels of abstraction. *Psychonomic Bulletin & Review*, *16*(1), 80–87.

Hubel, D. H., & Wiesel, T. N. (1979). Brain Mechanisms of Vision. *Scientific American*, *241*(3), 150–162. https://doi.org/10.1038/scientificamerican0979-150

Heron, J., Whitaker, D., & McGraw, P. V. (2004). Sensory uncertainty governs the extent of audio-visual interac- tion. Vision Research, 44(25), 2875–2884. https://doi.org/10.1016/j.visres.2004.07.001

Helmholtz H. 1867. Optique physiologique. Trans. É. Javal, T. Klein. Paris: Masson.

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal inference in multisensory perception. PLOS ONE, 2(9), Article e943. https://doi.org/10.1371/journal.pone.0000943

Koffka, K. (1922). Perception: an introduction to the Gestalt-Theorie. *Psychological bulletin*, *19*(10), 531.

Kinchla, R. A., & Wolfe, J. M. (1979). The order of visual processing:"Top-down,""bottom-up," or "middle-out". *Perception & psychophysics*, *25*, 225-231.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object Perception as Bayesian Inference. *Annual Review of Psychology*, *55*(1), 271–304. https://doi.org/10.1146/annurev.psych.55.090902.142005

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*(12), 712–719. https://doi.org/10.1016/j.tins.2004.10.007

Lewis, E. O. (1908). The effect of practice on the perception of the Müller-Lyer illusion. *British journal of psychology*, *2*(3), 294.

Løkke, H. (2008). The Stoics on sense perception. In *Theories of perception in medieval and early modern philosophy* (pp. 35-46). Dordrecht: Springer Netherlands.

Marr, D. (2010). *Vision: A computational investigation into the human representation and processing of visual information*. MIT press.

Michotte, A., Thinès, G., & Crabbé, G. (1991). Amodal completion of perceptual structures. *Michotte's experimental phenomenology of perception*, *1*, 140-167.

Moravec, L., & Beck, J. (1986). Amodal completion: Simplicity is not the explanation. *Bulletin of the Psychonomic Society*, *24*(4), 269-272.

O'Brien, D. (1970). The Effect of A Simile: Empedocles' Theories of Seeing and Breathing. *The Journal of Hellenic Studies*, *90*, 140–179. https://doi.org/10.2307/629759

Olivers, C. N. L., Meijer, F., & Theeuwes, J. (2006). Feature-based memory-driven attentional capture: Visual working memory content affects visual attention. Journal of Experimental Psychology: Human Perception and Performance, 32(5), 1243–1265. https://doi.org/10.1037/0096 -1523.32.5.1243

Ort, E., & Olivers, C. N. (2020). The capacity of multiple-target search. Visual Cognition, 28(5–8), 330–355.

Olivers, C. N. L., Peters, J., Houtkamp, R., & Roelfsema, P. R. (2011). Different states in visual working memory: When it guides attention and when it does not. Trends in Cognitive Sciences, 15(7), 327–334. https:// doi.org/10.1016/j.tics.2011.05.004

Palmer, S. E. (1999). *Vision science: Photons to phenomenology*. MIT press.
Pouget, A., Beck, J. M., Ma, W. J., & Latham, P. E. (2013). Probabilistic brains: Knowns and unknowns. *Nature Neuroscience*, *16*(9), 1170–1178. https://doi.org/10.1038/nn.3495

Parise, C. V. (2016). Crossmodal Correspondences: Standing Issues and Experimental Guidelines. *Multisensory Research*, *29*(1–3), 7–28. https://doi.org/10.1163/22134808-00002502

Rohe, T., & Noppeney, U. (2016). Distinct Computational Principles Govern Multisensory Integration in Primary Sensory and Association Cortices. *Current Biology*, *26*(4), 509–514. https://doi.org/10.1016/j.cub.2015.12.056

Sedley, D. (2018). Empedocles' theory of vision and Theophrastus' De Sensibus. In *Theophrastus* (pp. 20-31). Routledge.

Singh, M. (2004). Modal and amodal completion generate different shapes. *Psychological Science*, *15*(7), 454-459.

Sekuler, A. B., & Palmer, S. E. (1992). Perception of partly occluded objects: A microgenetic analysis. *Journal of Experimental Psychology: General*, *121*(1), 95–111. https://doi.org/10.1037/0096-3445.121.1.95

Soto, D., Heinke, D., Humphreys, G. W., & Blanco, M. J. (2005). Early, involuntary top-down guidance of attention from working memory. Journal of Experimental Psychology: Human Perception and Performance, 31(2), 248–261. https://doi.org/10.1037/0096-1523.31.2 .248

Soto, D., Hodsoll, J., Rotshtein, P., & Humphreys, G. W. (2008). Automatic guidance of attention from working memory. Trends in Cognitive Sciences, 12(9), 342–348. https://doi.org/10.1016/j.tics.2008.05 .007

Soto, D., & Humphreys, G. W. (2007). Automatic guidance of visual attention from verbal working memory. Journal of Experimental Psychology: Human Perception and Performance, 33(3), 730–737. https://doi.org/10 .1037/0096-1523.33.3.730

Soto, D., & Humphreys, G. W. (2008). Stressing the mind: The effect of cognitive load and articulatory suppression on attentional guidance from working memory. Perception & Psychophysics, 70(5), 924–934. https:// doi.org/10.3758/pp.70.5.924

Stryer, L. (1996). Vision: From photon to perception. *Proceedings of the National Academy of Sciences*, *93*(2), 557–559. https://doi.org/10.1073/pnas.93.2.557

Stokes, M. G., Atherton, K., Patai, E. Z., & Nobre, A. C. (2012). Long-term memory prepares neural activity for perception. *Proceedings of the National Academy of Sciences*, *109*(6). https://doi.org/10.1073/pnas.1108555108

Oliva, A., & Torralba, A. (2007). The role of context in object recognition. Trends in Cognitive Sciences, 11(12), 520–527. https://doi.org/10.1016/j.tics.2007.09.009

Palmer, Stephen, E. (1975). The effects of contextual scenes on the identification of objects. Memory & Cognition, 3(5), 519–526. https://doi.org/10.3758/BF03197524

Plass, J., Guzman-Martinez, E., Ortega, L., Suzuki, S., & Grabowecky, M. (2017). Automatic auditory disam- biguation of visual awareness. Attention, Perception, & Psychophysics, 79(7), 2055–2063. https://doi.org/10.3758/S13414-017-1355-0

Rohe, T., & Noppeney, U. (2015). Sensory reliability shapes perceptual inference via two mechanisms. Journal of Vision, 15(5), Article 22. https://doi.org/10.1167/15.5.22

Shams, L., Kamitani, Y., & Shimojo, S. (2000). What you see is what you hear. *Nature*, *408*(6814), 788–788. https://doi.org/10.1038/35048669

van Moorselaar, D., Theeuwes, J., & Olivers, C. N. L. (2014). In competition for the attentional template: Can multiple items within visual working memory guide attention? Journal of Experimental Psychology: Human Perception and Performance, 40(4), 1450–1464. https://doi.org/ 10.1037/a0036229

Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., & Von der Heydt, R. (2012). A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure–ground organization. *Psychological bulletin*, *138*(6), 1172.

Wagemans, J., Feldman, J., Gepshtein, S., Kimchi, R., Pomerantz, J. R., Van der Helm, P. A., & Van Leeuwen, C. (2012). A century of Gestalt psychology in visual perception: II. Conceptual and theoretical foundations. *Psychological bulletin*, *138*(6), 1218.

Watanabe, K., & Shimojo, S. (2001). When sound affects vision: Effects of auditory grouping on visual motion per- ception. Psychological Science, 12(2), 109–116. https:// doi.org/10.1111/1467-9280.00319

Wörgötter, F., Krüger, N., Pugeault, N., Calow, D., Lappe, M., Pauwels, K., Hulle, M. V., Tan, S., & Johnston, A. (2004). Early Cognitive Vision: Using Gestalt-Laws for Task-Dependent, Active Image-Processing. *Natural Computing*, *3*(3), 293–321. https://doi.org/10.1023/B:NACO.0000036817.38320.fe

Chapter 1 Guidance of Attention by Working Memory Is a Matter of Representational Fidelity

Jamal R. Williams, Timothy F. Brady, and Viola S. Störmer

As it appears in

Journal of Experimental Psychology: Human Perception and Performance

2022

Volume 48(3), Pages 202-231

AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology:
Human Perception and Performance

ISSN: 0096-1523

https://doi.org/10.1037/xhp0000985

Guidance of Attention by Working Memory Is a Matter of Representational Fidelity

Jamal R. Williams[1], Timothy F. Brady[1], and Viola S. Störmer[1, 2]
[1] Department of Psychology, University of California, San Diego
[2] Department of Psychological and Brain Sciences, Dartmouth College

Items that are held in visual working memory can guide attention toward matching features in the environment. Predominant theories propose that to guide attention, a memory item must be internally prioritized and given a special template status, which builds on the assumption that there are qualitatively distinct states in working memory. Here, we propose that no distinct states in working memory are necessary to explain why some items guide attention and others do not. Instead, we propose variations in attentional guidance arise because individual memories naturally vary in their representational fidelity, and only highly accurate memories automatically guide attention. Across a series of experiments and a simulation we show that (a) items in working memory vary naturally in representational fidelity; (b) attention is guided by all well-represented items, though frequently only one item is represented well enough to guide; and (c) no special working memory state for prioritized items is necessary to explain guidance. These findings challenge current models of attentional guidance and working memory and instead support a simpler account for how working memory and attention interact: Only the representational fidelity of memories, which varies naturally between items, determines whether and how strongly a memory representation guides attention.

Public Significance Statement
When you are holding an item in mind (say, your red mug), your visual attention is automatically guided toward red information in the environment. However, this does not always occur and seems to happen less often when you are holding in mind multiple pieces of information (say, your red mug and your blue coaster). This study demonstrates that the fidelity of a working memory representation alone may determine how strongly that item will interact with attention. Because memories vary randomly in fidelity and tend to be lower fidelity when holding more items in mind, this can explain why attention is sometimes, but not always, guided by items we hold in mind.

Keywords: attentional guidance, internal attention, variable precision, visual working memory

As we look around the world, we have the sense of a rich and complete perception. This is in spite of the fact that we are only able to process a small fraction of the available sensory information. To effectively and efficiently operate within this sensory maelstrom, some of that information must be prioritized: either because it is physically more salient or because it matches our current task goals and intentions. For example, when looking for a friend in a crowd, we may carefully hold their distinctive visual features in mind—like the red shirt they are wearing—as we try to find a match, and our attention will be guided toward matching features in the environment. But what if that memory representation is inaccurate or noisy, and your friend's shirt is in fact more orange than red? Would guidance occur in this situation? Surprisingly, although attentional guidance has been studied extensively, the relationship between the fidelity of a memory representation and how effectively that item can guide attention is not well understood. Instead, most work has focused on the number of items that can guide attention and whether such items must have a special status within working memory, like being in the "focus of attention." Here we propose that representational fidelity of memories alone—defined as how accurately an item is represented in working memory—is sufficient to explain why some items do and some items do not guide attention, independently of any special status of an item in memory.

Guidance by One Versus Two Items

It has been repeatedly found that when only a single feature is maintained in working memory, attention is automatically guided toward matching features in the environment (Olivers et al., 2006; Soto et al., 2005, 2008; Soto & Humphreys, 2007, 2008).

Jamal R. Williams https://orcid.org/0000-0002-3034-511X
Timothy F. Brady https://orcid.org/0000-0001-5924-5211
Correspondence concerning this article should be addressed to Jamal R. Williams, Department of Psychology, University of California, San Diego, 9500 Gilman Drive #0109, La Jolla, CA 92093, United States. Email: jrwilliams@ucsd.edu

However, it is less clear whether multiple working memory items can guide attention in a similarly incidental way. It is this, more incidental guidance, as opposed to a goal-directed guidance (e.g., Beck et al., 2012; Hollingworth & Hwang, 2013), that we focus on here and, at minimum, a significant literature suggests that guidance by multiple memory representations is more fragile than guidance by a single item, as often no incidental guidance effects have been observed at all when participants remembered more than a single item (Frătescu et al., 2019; van Moorselaar et al., 2014; see Ort & Olivers, 2020, for a review). Conversely, several studies report guidance when multiple items are held in mind, with some suggesting that attention is guided equally well by each memory item; Chen & Du, 2017; Hollingworth & Beck, 2016; Fan et al., 2019; Soto & Humphreys, 2008; (Zhang et al., 2011, 2018). Importantly, for guidance to occur, an item must be maintained in an active state within working memory (Olivers et al., 2006) and not simply primed (Kumar et al., 2009) or maintained for less-relevant, secondary tasks (Downing & Dodds, 2004).

To date, these differing results have mostly been discussed in terms of limits in visual working memory; typically, by focusing on the *number* of remembered items that can be prioritized by attention and thus given a special template status (Chen & Du, 2017; Fan et al., 2019; Frătescu et al., 2019; Hollingworth & Beck, 2016; Olivers et al., 2011; van Moorselaar et al., 2014; Zhang et al., 2018). To effectively search for an item in the environment, it has been proposed that we maintain a template: a representation in memory that resembles the item being searched (Olivers et al., 2011). Attentional template accounts presume that working memory is organized into qualitatively distinct states and propose that attention must be internally directed toward a memory representation—which elevates the item to the special, template status—for that item to interact with and bias attention (Chen & Du, 2017; Fan et al., 2019; Frătescu et al., 2019; Hollingworth & Beck, 2016; Olivers et al., 2011, 2011; van Moorselaar et al., 2014, 2014; Zhang et al., 2018). This emphasis on the template status as being the most important component of guidance leads to the commonly debated question of how many items can achieve this privileged status since, under this framework, any other (nonattended) items cannot guide attention (Hollingworth & Hwang, 2013; Olivers et al., 2011; van Moorselaar et al., 2014). This literature's focus on "how many items" is irrespective of whether guidance occurs automatically (e.g., van Moorselaar et al., 2014) or through top-down processes (e.g., Beck et al., 2012) and is consistent with the historically strong emphasis on quantifying visual working memory capacity by the *number* of representations that can be maintained (e.g., Cowan, 2001; Luck & Vogel, 1997).

Variation in Memory

A critical factor of attentional guidance, however, is that, for a remembered item to guide attention it must also be the case that it is an accurate and precise representation of the encoded item. Because attention is unlikely to be guided toward an item if the corresponding memory representation is weak, imprecise, or even focused on the wrong object, it is critical that we consider the quality of the memory representations themselves. All models of visual working memory capacity now acknowledge that items vary in precision, such that representational fidelity tends to be higher when only one item must be held in mind than for two items (e.g., Bays et al., 2009; Zhang & Luck, 2008). Additionally, many modern models see independent accumulation of noise across different items, and the resulting variation in fidelity across these items, as the core of visual working memory limits (Bays, 2015; Fougnie et al., 2012; Schurgin et al., 2020). Thus, even for very small set sizes, well within typical claims of three to four item "capacity limits" (Cowan, 2005), variation in how well an item accurately represents the originally encoded item (i.e., its representational fidelity) appears to be an inevitable byproduct of working memory storage that must be accounted for when considering how working memory is used to guide attention.

Variation in representational fidelity is also not solely about the overall decrease in average fidelity (memory strength) as more items must be held in mind. Individual items also vary within a single trial. For example, Fougnie et al. (2012) found that when participants remember three colors, they are far more accurate at reporting a color of their choosing compared with when they report the color of a randomly probed item (similar results are found by Adam et al., 2017). This finding could only occur if the fidelity of remembered items varied considerably within a trial. Why might items vary in representational fidelity within a trial? This variability has been proposed to arise from many sources, including differential encoding precision (van den Berg et al., 2012), how memory items relate to other items on the encoding display (Brady & Alvarez, 2015), differential prioritization through memory-related resource allocation (Bays & Husain, 2008; Bays & Taylor, 2018; Klyszejko et al., 2014), and differences in the representation of specific individual colors (e.g., Bae et al., 2015; Morey, 2011). Variation in fidelity has also been proposed to be a basic fact about the architecture of the working memory system, with noise corrupting items independently (Bays, 2015; Fougnie et al., 2012; Panichello et al., 2018; Schurgin et al., 2020; Wilken & Ma, 2004). These convergent results strongly suggest that variation between multiple working memory items is an inevitable and natural byproduct of working memory maintenance and is thus an important factor to consider for memory driven attentional guidance.

Can Attentional Guidance Be Explained by Memory Strength?

In the current work, we propose that differing results in the literature may be accounted for primarily by variation in representational fidelity. While some previous work has investigated how fidelity relates to guidance, such studies have often concluded that guidance occurs for template items (Dube & Al-Aidroos, 2019) and cannot occur for nontemplate items (Hollingworth & Hwang, 2013), irrespective of how well they might be represented in memory. However, in our view, these studies do not allow strong conclusions about the relationship between guidance and memory strength—an aggregate measure of representational fidelity—largely because of the way memory strength was measured. In particular, since nearly all current models of visual working memory acknowledge that single trial errors are stochastically related to the underlying representational fidelity of memory (e.g., Schneegans et al., 2020; Schurgin et al., 2020), such errors do not provide a strong basis for asking whether guidance is driven by variation in fidelity (see General Discussion for more). In contrast, in the current work, we directly test the relationship between memory strength and attentional guidance in a series of experiments that assess attentional guidance effects while precisely measuring memory strength of the remembered items. We ask participants to memorize two colors and use a simple search task that allows the detection of potentially small guidance effects. The implementation of a simple search task is in contrast to previous studies that have often used more complex search displays that may have taxed the already limited attentional and working memory systems, perhaps, inadvertently obscuring a multiple item effect[1] (Hollingworth & Hwang, 2013; Houtkamp & Roelfsema, 2006; Olivers et al., 2006; van Moorselaar et al., 2014; Woodman & Luck, 2007).

We hypothesize that, on average, representational fidelity—how well a memory item accurately represents the initially encoded item—is sufficient on its own to determine whether and how strongly that item will guide attention. That is, guidance does not depend on the number of items that can be prioritized within working memory but is determined by the representational fidelity of an item as well as the variation in representational fidelity between items that occurs naturally both across set sizes and within a single trial. We propose that the reason that guidance from multiple items is often found to be more fragile than single-item guidance is because it is less likely that multiple items are maintained in working memory with the strength and precision necessary to guide attention effectively.

The Current Work

Overall, across three main experiments and additional supplemental experiments (see Appendix), we find evidence to support an account where items vary naturally in their representational fidelity, and any and all memory items can guide attention insofar as they are well represented—even without a special template status. The critical and novel contribution of our study is based on a careful assessment of memory quality and its relationship with attentional guidance. Specifically, we precisely measure memory strength (d') which gives us a robust index of the overall memory quality across trials and allows us to infer the representational fidelity of memories on individual trials (i.e., how faithfully a particular item is represented on a single trial, which effectively represents a single draw from the memory strength distribution across trials). By probing participants' self-selected memories and comparing them to randomly selected memory representations, in Experiment 1, we show that memories vary in their representational fidelity naturally, and that typical multiple-item guidance effects are primarily driven by the most well-represented memory. In a series of simulations, we show that the observed variation between remembered items is expected and well characterized by a working memory model that directly predicts how much variation should be present (via signal detection theory; Schurgin et al., 2020). In Experiment 2, by experimentally manipulating the supposed "template status" of an item (with a retro-cue; van Moorselaar et al., 2014) and the representational fidelity of items (adding low or high perceptual-noise at encoding), we show that differences in representational fidelity can explain variation in attentional guidance, with internal attention simply being one of many ways to boost the fidelity of the attended memory. Finally, in Experiment 3 we show that attention is guided by well represented memories even when those memories do not achieve a special "template status" within working memory, finding guidance even for items placed outside "the focus of attention."

Our results can thus unify the seemingly irreconcilable findings that one or many working memory items can guide attention: When working memory resources are stretched among multiple active representations, often only a single item is represented well enough to guide (e.g., Exp. 1); however, in other cases, the representational fidelity of multiple items may be high enough to produce guidance from both items (Simulation, Exp 3, and Discussion). After presenting data from each experiment, we integrate across all of the experiments, and find compelling evidence that continuous variation in representational fidelity is sufficient to determine whether—and how strongly—an item guides attention, with no need to postulate distinct states in working memory, and that this accounts for guidance strength. Thus, these data support a model of memory-driven attentional guidance where representational fidelity fully explains how memories interact with attention and influence behavior. Overall, we conclude that (a) attention is guided by memories that accurately represent the encoded item, (b) that internal noise, accumulated independently for each item, determines whether an item is represented well enough to guide attention and (c) that no strict limits exist on the number of items capable of guiding attention; rather, any and all well-represented items can guide attention and that the magnitude of this guidance is directly related to an item's representational fidelity at that moment in time. We believe that this account explains and unifies many of the mixed results in the memory-driven attentional guidance literature and provides a new framework of how working memory and attention interact.

Experiment 1: Dissociating Guidance for Well-Represented and Poorly Represented Working Memory Items

Previous work has shown that memory driven attentional guidance is less strong when two items are maintained in working memory compared with one (e.g., van Moorselaar et al., 2014) and that representational fidelity varies considerably between actively maintained items in working memory tasks (e.g., Fougnie et al., 2012). We replicate these findings using our own paradigm in the Appendix, finding in Experiment A1 that guidance is less strong when two items are maintained than one; and finding in Experiment A2 that representational fidelity differs across memory items, such that 1 item tends to be maintained more accurately than the other when two items must be held in mind in an attentional guidance task (see Appendix).

In Experiment 1, we investigated whether this variance between multiple items in representational fidelity can explain the differences in guidance strength on any particular trial. In short, we asked whether the best represented item is generally responsible for the guidance effect by having participants perform both a search task and memory task on each trial. Participants maintain two unique colors in visual working memory for a memory task that occurred at the end of each trial. Prior to reporting one of the remembered colors on a continuous color wheel, participants performed a

[1] We have, however, expanded search displays to contain four items and observed guidance when multiple items are held in mind and the search display contains four items (see Appendix).

visual search task in which color was irrelevant. At the end of each trial, either one of the remembered items was randomly probed (forced report) or participants would freely report one item of their choosing (free report). Evidence suggests that people are nearly optimal at choosing the more precise item for report (Fougnie et al., 2012) and we hypothesized that, even without any explicit experimental manipulation—participants were simply instructed to report any memory item that they wished—individuals would have knowledge about the representational fidelity of each item and, if some asymmetry between the items exists, would select the more precise item to report. Thus, free report trials allow us to estimate the representational fidelity of selected items, compare it to randomly probed items, and examine whether items differ in their representational fidelity without explicit manipulation.

The primary goal of Experiment 1 is to relate the representational fidelity of individual items to the guidance effects during search. Specifically, by focusing on free report trials, we can sort trials based on whether a chosen memory item happened to be present in the search display that occurs before the memory report (chosen-item: present) or not (chosen-item: absent). Our critical condition here is chosen-item: absent, where we expect to find a diminished guidance effect (that could even be zero) because we expect items with less-than-optimal representational fidelity to exert limited guidance over attention. Chosen-item: absent trials should on average represent the amount of guidance that occurs from poorly represented items since it is unlikely that participants would choose a poorly represented item to report when they have multiple items to choose from (Fougnie et al., 2012). By contrast, chosen-item: present trials likely represent a mixture of items that were well represented from the initial encoding (those that should guide) and trials where items are less well represented during the search task, but that are nonetheless chosen during the memory report, perhaps of the brief reexposure to that color during visual search.[2] Nonetheless, any difference in guidance between these two trial types suggests that representational fidelity differences between the items are related to differences in attentional guidance.

Method

The design, sample size, exclusion criteria, and analysis plan for this experiment were preregistered using AsPredicted (http://aspredicted.org/blind.php?x=7b5y74).

Participants

Consistent with our preregistration, the final sample included thirty participants (24 women) from UC San Diego, who took part in this study in exchange for course credit. Our primary question of interest was whether guidance would be different on free report trials where the item from the search display was chosen than trials where the item was not chosen. Pilot data suggested an effect with Cohen's $d_z > 0.5$. Thus, per our preregistration, we determined that 30 participants would provide adequate power (power = .8) to detect effects of Cohen's $d_z = 0.5$ at an alpha level of .05 using the pwr.t.test in *R* (all subsequent power calculations used this same package and were for the same power and alpha levels). Data from four additional participants were removed and replaced for failing to meet the preregistered exclusion criterion and, as in the appendix experiments, data from another participant was removed and replaced for failing to follow forced-report instructions and thus failing to report the probed memory item on more than 40% of trials.

Stimuli

Stimuli were generated and presented using MATLAB and the Psychophysics Toolbox (Brainard & Vision, 1997; Pelli & Vision, 1997). Memory items were colored rings that were 3° visual angle in diameter, .3° thick, and were centrally placed 4° to the left or right of fixation. On every trial, the color of one memory item was randomly drawn from a uniformly spaced circle (radius 49°) extracted from the CIE L^*a^*b space, centered at ($L = 54$, $a = 21.5$, $b = 11.5$) and the second color was selected to be 90° away in color space from the first color (with ±5° of jitter). The search display consisted of a target line which was .3° thick, .4° long, tilted .06° to the left or right of vertical, and placed 4° above or below fixation and a single vertical distractor line that was placed at the opposite location (see Figure 1). The target and distractor lines were encircled in colored rings that matched one of the memory item properties. One of the colors matched a memory color and the other color was chosen to be 180° away from it in color space (this was 90° away from the other memory item). On the random probe memory display, one of the memory items initially appeared in gray (identical features to memory items) surrounded by a continuous color wheel which was 15° in diameter, .3° thick, and was centrally placed about fixation. On free report displays, two gray placeholders were presented and after one of these placeholders was selected, the continuous color wheel appeared.

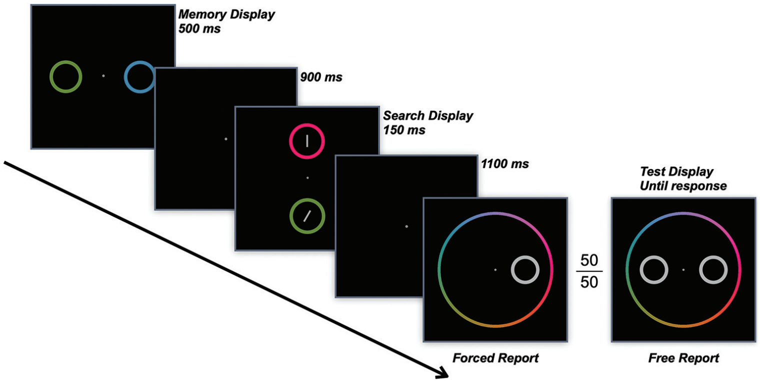
Procedure

Each trial started with the presentation of two memory items (500 ms) that were to be remembered for a memory task at the end of the trial. Following a 900 ms delay, a search display appeared for 150 ms. The search display contained two lines, a distractor and a target line, above and below fixation. Participants needed to rapidly determine the orientation of the tilted, target line after the search display disappeared. Feedback to respond more quickly was provided when responses exceeded 1,200 ms. After participants provided a search response, and after a delay of 1,100 ms participants were probed on one of the memory items. On forced report trials a randomly selected memory item was presented in gray and would change color as participants moved the mouse around the continuous color wheel. On free report trials, participants were presented with two gray rings to the left and right of fixation. Here, they clicked on which item they would like to report prior to using the continuous color wheel (see Figure 1).

The memory task was evenly divided between the free report and forced report conditions and on half of all forced report trials the probed item was present in the visual search display and was

[2] Note that because participants were required to select at least one item to report, it is feasible to imagine that on some trials both items would have less than optimal fidelity and would produce a weakened guidance effect. Furthermore, although the search display is presented briefly so as to discourage intentional re-encoding, we have also replicated the same pattern of data in another experiment in which we used two set sizes (1 vs. 2) along with the free-report method, but had participants perform either search or the memory task on separate trials to avoid the possibility of re-encoding items during the search (see Appendix).

Figure 1
Experiment 1 Task Design



Note. Participants remember two colors on every trial then perform a visual search task after a short delay, followed by a memory report at the end of each trial. On half of the trials participants are forced to report a particular item (left memory example), and on the remaining half of trials they can choose which color they want to report (free-report trials; right). This allows us to separately analyze the search benefit for free-report and random-probe memory conditions. See the online article for the color version of this figure.

absent on the remaining half. At test, participants were asked to use the mouse to find the color closest to the remembered color on the color wheel. The location of the test-item indicated which memory item should be reported (e.g., a test-item on the left probed the color of the memory item that was on the left at encoding), and which item was tested was counterbalanced across the experiment. Once the mouse was moved from the central fixation point the gray test-item changed color to match the color at the position of the mouse cursor. Once participants identified the color that matched the remembered color as precisely as possible on the color wheel, they locked their response by clicking the mouse button. Response error, defined as the difference in degrees between the provided response and the correct answer, was shown after every memory trial and participants were instructed to keep this error below 10°. Participants were instructed to prioritize speed without compromising accuracy for the search task and, for the memory task, were instructed to prioritize precision without compromising temporal efficiency.

Analysis

In the search task, we calculated each participant's median response time (RT) where the target was in the memory-matched color and in the distractor color separately. Our main measure of interest is in the magnitude of the difference between these match conditions and will refer to this RT difference as the amount of guidance. Note that this measure indexes both benefits of being faster to the target-match trials as well as costs of being slower to the distractor-match trials. Here, we are agnostic to these differential effects and thus simply summarize them in the RT difference (Raw RTs per condition are shown in the tables in the Appendix). RTs that were faster than 200 ms or slower than 1,500 ms were removed prior to any further analysis. All subsequent analyses use these criteria, unless otherwise noted.

Memory performance was evaluated in two ways. First, we used a simple descriptive statistic, angular deviation (a circular analog of standard deviation) to provide a nonparametric estimate of an item's representational fidelity. Second, we implemented the Target Confusability Competition (TCC) model to better quantify memory strength (Schurgin et al., 2020). The TCC model is based on recent evidence showing that continuous report memory distributions can be quantified by a single parameter — memory strength (i.e., d')—once the nonlinear nature of perceptual similarity is accounted for (Schurgin et al., 2020). All statistical analyses on memory are performed using memory strength (d'). Specifically, for any given color wheel, there is a completely fixed perceptual similarity function that quantifies how confusable colors are, but this function is not linearly related to distance along the color wheel—rather, it is roughly exponential (as in Fechner's law).

Understanding this confusability function allows a simple signal detection model to explain working memory performance across a huge variety of conditions, with only a single parameter (d'). In particular, on any given trial, the to-be-remembered color is boosted by a strong familiarity signal (strength: d'), and completely dissimilar colors do not have their familiarity signal boosted at all. Intermediate colors have their familiarity signals boosted proportional to how similar they are to the target. So, a color 1° away from the to-be-remembered color gets a large boost in familiarity, and a color 10° away from the to-be-remembered color gets a moderate boost in familiarity. Noise is then added to these familiarity signals, and when participants are asked what color they saw, they report the color that has the highest familiarity.

Formally, this means the continuous report task is conceptualized as a 360-alternative forced choice task: Let $f(x)$ be how similar a given color is to the to-be-remembered color. Let (X_{-179}, . . .,

X_{180}) be a vector of normal random values with means $d_x = d' f(x)$ and unit variance. Then the reported color value, *r,* on this trial is simply:

$$r \sim \text{argmax}(X_{-179}, \ \ldots, \ X_{180})$$

In the current data, which uses the same color wheel used by Schurgin et al. (2020), we rely on their similarity data and their technique for fitting the model, including the necessary correlation between colors based on perceptual matching data to adjust d' (Schurgin et al., 2020).

Since the current study introduced a larger possibility of location confusions than the data fit by Schurgin et al. (2020), we introduce a "swap" parameter into the model (as in Bays et al., 2009). We report memory strength (d'), the proportion of trials where the nonprobed item was incorrectly reported (swap rate), and an adjusted d' which conservatively assumes participants had no information about the probed item. In particular, rather than assuming participants always report based on the similarity, $f(x)$, to the target color, we assume that on some trials, participants instead respond based on similarity to the nontarget (i.e., the item at the nontested location). Let $f(x)$ be the similarity to the target color and $g(x)$ be the similarity to the nontarget. Let (X_{-179}, . . ., X_{180}) be a normal random vector with means $d_x = d' f(x)$ and unit variance, and (Y_{-179}, . . ., Y_{180}) be a normal random vector with means $d_x = d' \ g(x),$ and unit variance. Let β be the "swap rate". Responses are generated as follows:

$$\begin{aligned} &w \sim \text{Bernoulli}(\beta) \\ &r \sim w \times \text{argmax}(Y_{-179}, \ \ldots, \ Y_{180}) \\ &\quad + (1-w)\text{argmax}(X_{-179}, \ \ldots, \ X_{180}) \end{aligned}$$

In other words, for each trial, participants report the maximum familiarity signal from either the target or nontarget with some probability β of reporting from the nontarget distribution. We again make use of the Schurgin et al. (2020) similarity data and perceptual matching data.

Whereas for conditions with no swaps (e.g., single-item condition), d' alone provides a measure of how strong the memory was, in the presence of swaps it is unclear what to consider as the strength of the underlying memory. The d' parameter in the swap model is best thought of as "how strong memory would be if people always reported based on the correct target" (e.g., ignoring any contribution of swaps). This value thus only represents the actual memory strength across all trials if "swaps" occur totally based on response error, and even when participants misreport items, they always have a very strong representation of the correct target as well. By contrast, if participants tend to report an incorrect item as a form of "strategic guessing," for example, selectively do so when they have very little information about the correct item (Pratte, 2019)—then the best way to understand memory strength across all trials is to reduce the estimated d' by the proportion of swaps (e.g., assume on "swap" trials, people had $d' = \sim 0$ for the correct item). Telling apart these accounts—or where on this continuum people tend to be—is difficult and not our main purpose, and d' is an approximately interval measure, unlike percent correct or "guess rate" (Macmillan & Creelman, 2005). Thus, for our results, we report three measures of memory strength: angular deviation (just a descriptive statistic of how tightly clustered participants error are); d' (memory strength on trials where the correct item was reported); and adjusted d' (memory strength after adjusting downward to account for the possibility of no memories when participants made location swaps, e.g., d' [1 − β]). In general, all of our conclusions hold similarly for angular deviation and both d' measures and thus hold regardless of what assumption is made about memory for the target on swap trials.

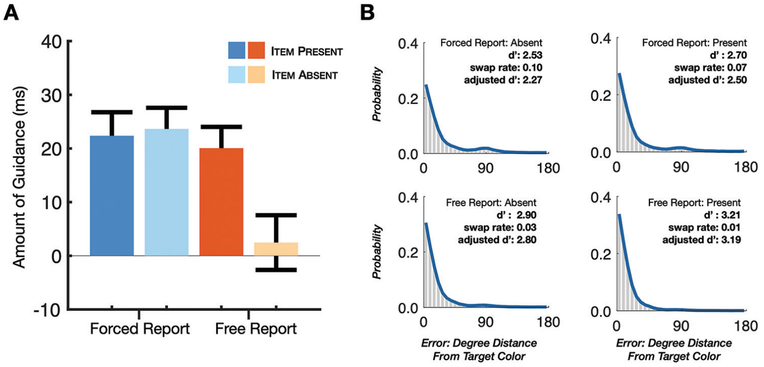
Results and Discussion

Our main analysis focused on the magnitude of attentional guidance, operationalized as the difference in RT between target- and distractor-match trials, separately for each condition. We first submitted the average guidance effect, separated by memory condition (forced or free report) and search type (memory-item: present or absent), to a repeated measures ANOVA. This analysis revealed a significant main effect of memory condition, $F(1, 29) = 4.19$ $p =$ .04, no main effect of search type, $F(1, 29) = 2.09$, $p = .16$, and a significant interaction, $F(1, 29) = 7.73$, $p = .009$. Next, we examined the amount of guidance on forced report trials alone and found that participants were faster on target match compared with distractor match trials regardless of whether the search display contained an item that was subsequently tested, $t(29) = 3.96$, $p <$ .001, $d_z = 0.81$, or not ($t(29) = 4.46$, $p < .001$, $d_z = 0.73$, and that the amount of guidance was consistent across these conditions as expected, $t(29) = .21$, $p = .83$, $d_z = 0.08$. To determine whether we have evidence to accept the null effect, we used Bayes Factors with a standard scale of the effect size (.707) and the Jeffrey-Zellner-Siow Prior (JZS; Rouder et al., 2009). Here, we found that there is strong evidence to support no difference between these conditions ($BF_{01} = 5.04$).

Next, we examined the free-report condition alone. For chosen-item: present trials, we found that participants were significantly faster on trials where the chosen working memory item happened to match the color that surrounded the search target compared with distractor-match trials, $t(29) = 3.6$, $p < .001$, $d_z = 0.66$; see Figure 2. On the critical condition, chosen-item: absent, we tested whether unselected memory items would guide attention. Here, we found no significant difference between target- and distractor-match trials, $t(29) = .37$, $p = .72$, $d_z = 0.06$, found evidence to support the null finding ($BF_{01} = 4.84$), and found a significantly larger amount of guidance on chosen-item: present trials compared with absent trials, $t(29) = 2.48$, $p = .01$, $d_z = 0.59$. Thus, the item that was not chosen in the memory task had little observable influence on visual search performance.

We next focused on memory performance to understand the relationship between guidance and memory strength. We submitted memory performance (swap d') to a repeated measures ANOVA. The main effects of report condition (forced or free report; $F(1, 29) = 73.99$, $p < .0001$), and search type (whether a tested memory item appeared in the search display; $F(1, 29) =$ 18.15, $p < .001$) were significant; with no significant interactions (Fs < 1.61). We next compared memory strength (d') across conditions and report all memory measures comprehensively. Memory performance was overall quite good and was better on free-report trials (TCC $d' = 3.05$, *swap rate* = .02, *adjusted* $d' = 3.00$, $SD = 28.36°$) compared with random-probe trials (TCC $d' = 2.62$,

Figure 2
Experiment 1 Results



Note. (A) Amount of guidance (RT for distractor minus target match trials) for forced report (blue bars; left pair) and free report (orange bars; right pair) trials separately for when the reported memory item was present or absent in the search display (dark vs light colors). For forced report trials, there is a clear guidance effect when the subsequently probed memory item was or was not present in the search display. For free report trials, by contrast, where participants are able to selectively report their strongest memory, there is a larger guidance effect when the subsequently chosen free report item was present in the search display than when the chosen item was absent. Thus, a memory item that is not subsequently chosen for report in free report exerts little to no influence over attention; resulting in a much smaller difference in RT for target and distractor match trials than a chosen item. (B) Memory performance for forced report (top row) and free report (bottom row) separated by whether the reported item appeared in the search display or not. Memory strength was better on free report compared with forced report trials, suggesting participants report their strongest memories. In both cases, memory was slightly stronger when the relevant memory item was seen again on the search portion before the memory probe. Overall, then, memory was best when the freely chosen item was present on the previously encountered visual search display. See the online article for the color version of this figure.

swap rate = .08, *adjusted* d' = 2.38, *SD* = 38.7°; TCC d' = 3.05 vs 2.62, respectively; $t(29) = 7.37$, $p < .001$), as expected. Memory performance was superior when the freely reported item was present in the previously seen visual search task (TCC d' = 3.21, swap rate = .01, adjusted d' = 3.19, *SD* = 26.25°) compared with when it was not (TCC d' = 2.90, swap rate = .03, adjusted d' = 2.80, *SD* = 30.47°; $t(29) = 3.36$ $p = .01$), suggesting a benefit arose from reexposure. To check whether participants were biased to choose the memory item that was briefly presented during visual search more often than the absent item, we compared the proportion of chosen-items: present to chosen-item: absent trials. We found that participants reported the item that was absent from the search display about as often (47% of trials) as items that were present; which suggests that participants were not more likely to report an item that was present in the search display. Furthermore, when participants reported an absent item, they tended to have a very strong representation of it (d' = 2.90) and rarely reported the wrong item (*swap rate* = .03). Thus, participants were not biased to select an item that they had previously seen in the search display despite being briefly reexposed (for 150 ms) to that color.

Overall, in Experiment 1, we find that one item tends to be better represented than another, as free-report memory probes result in higher fidelity memories, and that less well-represented items are unlikely to guide attention. In particular, we find little to no evidence of attentional guidance by a memory representation that is not chosen by participants as their strongest memory. Additionally, when two items are maintained in working memory, we find that guidance can largely be explained by a single working memory item (similar to Beck & Vickery, 2019; van Moorselaar et al., 2014). These results demonstrate that multiple items are represented with varying levels of representational fidelity and appear to exert correspondingly differential influences over attentional guidance.

Simulation: Representational Fidelity Naturally Varies Between Items Even When All Items Are Encoded Equally Well

Experiment 1 revealed that working memory items naturally vary in how accurately participants can report them: Allowing participants to freely choose a memory item to report results in improved memory performance for the chosen item relative to a randomly probed item. Furthermore, the chosen item—the one with better memory performance—primarily guides attention, whereas the other, less precise item (that is not chosen) has little influence on visual search. Although this shows a link between guidance and memory strength, it does not provide any evidence against a special template or focus of attention account. In particular, while the majority of models suppose stochasticity is sufficient to explain why memory strength varies (e.g., Fougnie et al., 2012;

Schurgin et al., 2020; Schneegans et al., 2020), another possible explanation for this variation in memory strength is that it too is caused by directing attention to one of the items—either externally, during encoding or internally, during the delay period, thereby giving some items a special status within the structure of working memory (e.g., Oberauer & Lin, 2017). Such a model is based on the idea that working memory is divided into qualitatively distinct states, and that item(s) can achieve a special status, which results in a strong memory for that item and strong guidance during visual search (Olivers et al., 2011). Such an account is possible and consistent with the data so far but this model makes strong assumptions, namely that working memory consists of fundamentally different memory states.

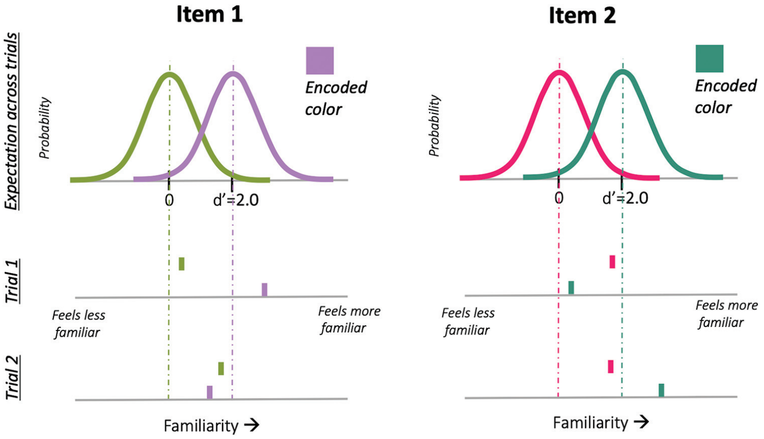
In many ways, a more simple possibility is that items intrinsically vary in how accurately they reflect the encoded item, and that items that are poorly represented simply cannot effectively guide attention. Such between-item variation appears to arise naturally due to noise in memory that is independent for each item (e.g., Fougnie et al., 2012; Panichello et al., 2018; Schurgin et al., 2020). For example, Fougnie et al. (2012) found that while items vary in precision, this variation is not at all correlated across items—contrary to what you'd expect if there is some overarching attentional resource that is being unevenly distributed across items (as suggested by attentional template accounts).

In a series of simulations, we answer two related questions: Is the amount of variation we find between the two items in memory reports consistent with this natural variation account? Or does it require additional assumptions about special attentional states within working memory?

Importantly, variation between items doesn't need to be explicitly accounted for in the TCC model of Schurgin et al. (2020) that we use to fit memory distributions throughout the current article, and despite this, this model makes precise predictions about how memory reports should vary between items, but it predicts this without explicitly modeling variation between items. Interitem variability in this model is simply as a natural consequence of a signal detection process (i.e., independent accumulation of noise for each item; see Figure 3). The expectation of noise or variability in the absence of an overarching attentional resource is common in models of memory and several models of working memory make precise and explicit quantitative predictions about how much variation we should expect between items at a given set size. For example, van den Berg et al. (2012) propose a particular mathematical relationship between variability and set size, with

Figure 3

A Signal Detection Account of Memory



Note. Imagine a scenario where people encode two items, one in one location and the other in a separate location, and have independent memories for the two items. In this case, both items have the same underlying memory strength, in the sense that the signal to noise ratio for both is $d' = 2.0$ on average. That is, the encoded color is "boosted" in familiarity by two standard deviations ($d' = 2.0$) relative to how familiar an unseen, completely distinct color is (which is centered at familiarity = 0). Thus, on average, participants find the color they actually saw at each location the most familiar. However, on any individual trial, the representation of one item may end up more or less accurately reflecting the original color due to independent noise. That is, signal detection conceives of each color for each item as varying in familiarity trial-to-trial (e.g., any given trial is a sample from the across trial distribution). Thus, the representational fidelity of an item varies from trial to trial from this noise process, even with the same underlying memory strength (d') for both items. In signal detection, confidence arises from the same familiarity signal as the decision. Thus, whichever item ends up the strongest "winning" signal (e.g., the strength of the most familiar color) is the one we'd expect people to report in a free report task. See the online article for the color version of this figure.

variability following a gamma distribution within set size and the mean of this distribution varying according to a power law across set sizes.

TCC, on the other hand, like all signal detection-based accounts of memory, proposes that the familiarity signals we use to decide which item we saw are inherently noisy. That is, even if you never saw a green item on a given trial, green might feel more familiar or less familiar (the top left of Figure 3 shows the probability distribution across trials of how likely green is to feel each level of familiarity). Although actually seen items are on average more familiar than items you haven't seen (as reflected in a higher familiarity, on average shifted by $d' = 2.0$, of the purple item in Figure 3), they also vary in familiarity, such that they might feel more familiar (purple item on trial 1) or less familiar (purple item on trial 2) across trials.

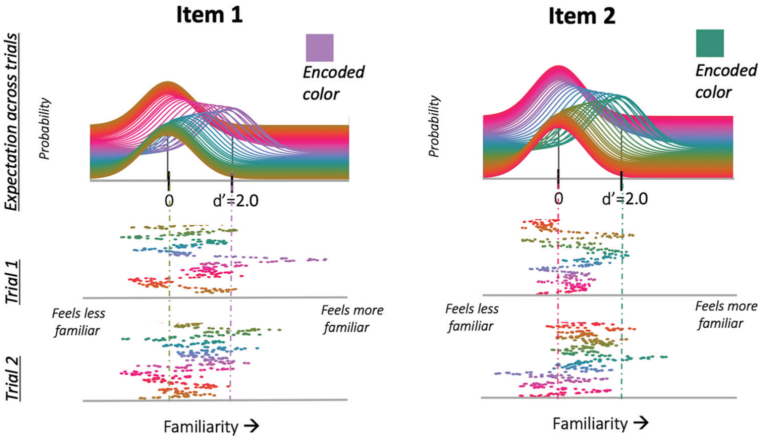
Similarly, within a given trial, the TCC model proposes that item representations accumulate independent noise (Schurgin et al., 2020; see also, Fougnie et al., 2012). So, if you encode both a purple item and a green item, then on a particular trial, the purple item may happen to feel familiar, and the green item might feel less familiar. Importantly, simply knowing the underlying memory strength (d') of the items *on average* allows us to predict exactly how much they should vary *trial to trial* —because d' is a signal-to-noise ratio, we can use it to infer exactly how much variation in ultimate memory performance there should be between items that accumulate independent noise. The TCC model makes a slightly more complex prediction than the simple signal detection theory account, because this model predicts not just how likely people are to endorse having previously seen purple, but also exactly how likely they are to endorse any other color as the previously seen color. However, this same signal detection-based logic applies equally to this model, with the added idea that when you encode purple, not only does purple feel more familiar, but all similar colors also get enhanced familiarity.

In sum, this model says that while the noise is independent across different items, within the representation of an item, the familiarity signal is not independent for each of the colors but depends on the underlying perceptual similarity structure: if purple is encoded, not only does purple get a boost in familiarity (d'), but similar colors (e.g., pink) get a boost, more so than distinct colors (e.g., yellow; see Figure 4 and Method from Exp. 1). When we add noise, this makes participants more or less likely to endorse particular colors as the most familiar, and this differentially impacts different items.

How does this a priori prediction of how much items should vary in memory performance relate to the actual variability observed in the free report condition of Experiment 1? To test this, we take the d' estimated from the overall average performance across items on forced report trials and use this to simulate (a) how much variability we expect between items in terms of performance and (b) how this relates to actual free report performance. In particular, we assume that during initial encoding, all

Figure 4

In the Target Confusability Competition Model (TCC), the Familiarity Signal for Particular Colors Depends on the Fixed Underlying Similarity Structure of the Color Representations



Note. When one color is encoded, not only that particular color gets a boost in familiarity (of d'), but colors that are perceptually similar to that color are also enhanced and thus more likely to feel familiar relative to distinct colors. Added noise on individual trials results in differential representational fidelity across items within a trial, even when each item has the same initial memory strength (d'). For example, on Trial 1, this individual has a lot of confidence that the purple item is some kind of purple, whereas the most activated color channel for the green item is yellow, so on this trial, the representational fidelity of Item 1 is greater than that of Item 2, even though both have the same underlying signal-to-noise ratio (d'). See the online article for the color version of this figure.

memory items are encoded equally well (with the same d'), resulting in the same familiarity boost of the to-be-remembered color and similar colors (see Figure 4). Then, during the delay period, noise is added to all color representations, separately for each item, which changes the familiarity signals for each of these color representations.

In a standard forced probe situation, participants report the color they find to be most familiar for the probed item. In the free report condition, they consider the most familiar color from both items, and choose to report the color for the item that has a higher familiarity. We can then ask whether the amount of variation we observe in our data—for example, the improvement of memory reports in free report trials relative to forced choice trials—is consistent with the variation predicted by this framework, or whether it exceeds it and thus calls for another explanation (like a special attentional focus that biases item representations systematically).

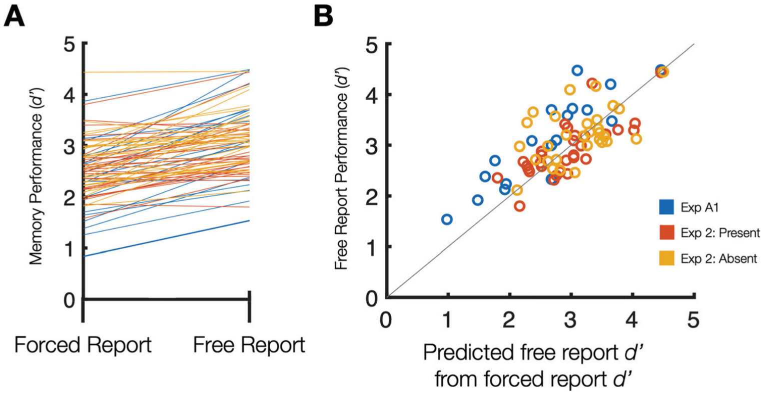
Figure 5A shows that across both Experiment 1 and Experiment A2 from the Appendix, free report results in reliably higher memory performance than forced report. As shown in Figure 5B, we find that this difference between free report and forced report matches the prediction that both items were encoded with the same signal and had a similar amount of noise added to them (r = .73, $p < .0001$). Thus, the variability between items that necessarily arises from the signal detection process is sufficient to explain the variability in memory performance that we observe. This demonstrates that assuming intrinsic variation in the representational fidelity of memories—attributable to independent item noise—is sufficient to explain the difference in memory performance we observed in our previous experiment. This assumption is also consistent with data indicating items vary in precision independently of each other (Fougnie et al., 2012). Overall, our simulations argue that attentional guidance is driven by an item that happens to have a stronger representational fidelity—and most accurately represents the initially encoded color—on that trial simply as a result of natural variation, and without needing to confer any special status on an item.

Experiment 2: Effects of Attentional Cues and Representational Fidelity on Search

Our data, together with the simulations, are consistent with a representational fidelity account. In particular, memory representations vary naturally due to noise, producing asymmetries in representational fidelity between items, and that guidance effects are present when the item that happens to have a strong memory representation on a given trial is present in the search display. This explanation differs from attentional template frameworks which embrace distinct states among working memory items and propose that attention is exclusively guided by any item that achieves a special and prioritized template status. Critically, such theories maintain that the template status is the single most important factor to predicting guidance and report that the precision of template item(s) in memory has little (Frătescu et al., 2019; van Moorselaar et al., 2014) to no impact on attentional guidance (Zhang et al., 2018; but see Dube & Al-Aidroos, 2019; Fan et al., 2019; Hollingworth & Beck, 2016; Hollingworth & Hwang, 2013; Hout & Goldinger, 2015; Rajsic et al., 2017).

For example, Dube and Al-Aidroos (2019) found that a 100% valid attentional retro-cue resulted in attentional guidance, but that a 70% valid cue did not; despite producing indistinguishable

Figure 5
Simulation Results



Note. (A) Individual lines are subjects: This shows that free report reliably exceeds forced report across all experiments. (B) Signal-detection based prediction about free report with the strong assumption that all of the variability comes simply from the independent noise added to each item: e.g., with the assumption that d' for both items is exactly what is estimated from forced report (and thus the same initially), and the only difference in free report is that people report their most confident memory (e.g., the signal that has the strongest familiarity on that trial). This provides an excellent explanation of free report performance, with no actual difference between the items other than the noise process predicted by signal detection. See the online article for the color version of this figure.

memory performance between the two conditions. They concluded that memory strength alone is not sufficient to grant a memory item with the template status and thus guide attention. Similarly, Hollingworth and Hwang (2013) report that an uncued item does not guide attention irrespective of being very well-represented and suggested that this is because the item lacked a template status (a finding which is contradicted by Zhang et al., 2018). Yet, here, we have shown that memory strength is highly predictive of search by demonstrating that well-represented memories guide attention and that less well-represented items do not (Exp. 1), and that the focus of attention is not necessary to explain the observed variation in fidelity between free and forced report (Simulation). Our results thus far are consistent with the proposition that representational fidelity primarily determines whether and the extent to which an item will guide attention: as a representation becomes less and less identical to the initially encoded item, it will exert guidance over attention that is equally less and less efficacious.

In Experiment 2, we asked whether memory performance (d', an aggregate measure of representational fidelity), or a special focus of attention better predicts guidance effects. In most studies, including our Experiment 1, it has been difficult to estimate representational fidelity and attention separately since attended items are usually maintained more precisely than other working memory items (see Oberauer & Lin, 2017; Rajsic et al., 2017). Thus, we next separated the influence of representational fidelity and attentional focus on guidance by independently varying attentional focus and representational fidelity across trials. Participants performed a similar task to before, except now, to manipulate representational fidelity, we added different amounts of perceptual noise during encoding. This has been shown to increase confusability between colors, thus decreasing the signal to noise ratio of the memory representations without manipulating attention (see Zhang & Luck, 2008). We also varied attentional focus using a directional retro-cue which has been shown to change the attentional state and improve the representational fidelity of an item by protecting it from interference (e.g., Oberauer, 2002; Oberauer & Lin, 2017).[3]

The goal of Experiment 2 is thus to change the representational fidelity of memory items without changing the focus of attention, and vice versa, and to test whether differences in fidelity can affect visual search performance, independently of an item's attentional status. If fidelity plays an important role, guidance should vary according to how well an item is represented. And, if attention is simply one way of many to modulate memory strength, we'd expect guidance to be greater for attended items but still vary depending on the representational fidelity at the time of search. If, however, the fidelity of the remembered item plays little to no role, as previously stated by attentional template accounts (Dube & Al-Aidroos, 2019; Fan et al., 2019; Hollingworth & Beck, 2016; Hollingworth & Hwang, 2013; Rajsic et al., 2017; Zhang et al., 2018), once an item has achieved template status by being focally attended, then we should find a similar sized guidance effect across all items with a template status, regardless of how well they are represented.

Method

The design, sample size, exclusion criteria, and analysis plan for this experiment were preregistered using AsPredicted (https://aspredicted.org/blind.php?x=xx5wr8).

Participants

The final sample included 50 undergraduates (34 women, mean age = 22.47 y) from UC San Diego who took part in this study in exchange for course credit. Because we assumed the external noise manipulation would result in a smaller effect than the effect of internal noise (which was ∼.65), the preregistered sample size of 50 was powered to allow us to detect effects considerably smaller than those observed in Experiment 1 ($d_z = 0.4$). Data from an additional eight participants were removed for failing to meet the predefined inclusion criteria and, as in the previous experiments, data from three others were removed for incorrectly reporting the probed memory item on at least 40% of trials (greater than 3 standard deviations from the group average).

Stimuli

Memory items were presented as squares with a side length of 3° in visual angle. On high-perceptual-noise (high-noise) trials, 360 uniquely colored dots (each .23° in diameter) were randomly positioned within an invisible circle (6° in diameter) over each memory item. On low-perceptual-noise (low-noise) trials, 60 uniquely colored dots were randomly positioned over both memory items. To achieve an even distribution of colors within the perceptual noise, the colors of the dots were chosen to be 1° and 6° apart in color space for high- and low-noise trials respectively. Postcue displays consisted of either a neutral-cue (two arrows, each facing away from fixation and toward the memory items) or a directional-cue (one arrow facing one of the memory items; each arrow was .64° long and .1° thick). The memory report display was identical to previous experiments except that the test-item(s) were squares with equal proportions as the memory items.

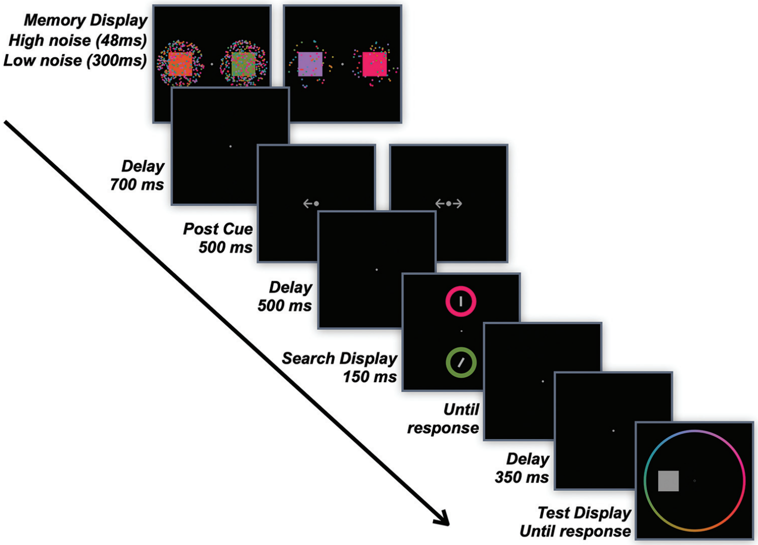
Procedure

High-noise displays were presented for 48 ms while low-noise displays were presented for 300 ms. Following the presentation of the memory items, there was a 700-ms delay. After this delay, on 256 trials for both high- and low-noise trials, participants were shown a neutral postcue; on the other 256 trials, they were shown a directional postcue which cued them to the item that was to be probed in the final memory task with 100% validity. After a further delay of 500 ms, participants performed the visual search task. After the search task—briefly flashed for 150 ms and followed by an untimed response window—and a further delay of 350 ms, the memory report display was presented. On neutral cue trials, there was a 50% chance of each item being probed and on a directional cue trial the cued item was always tested. On both kinds of trials, 50% of the time the subsequently tested item was the memory-matched color from the visual search display and on 50% of these trials, the memory-match color was the visual search target, and on 50% of these trials the memory-match item was the distractor (see Figure 6).

[3] Note that although some purport that the retro-cue facilitates memory performance by giving it a special status (i.e., placing it within the focus of attention), it is equally likely that retro-cue effects arise from the flexible allocation of a continuous memory resource (see Bays & Taylor, 2018).

Figure 6

Experiment 2 Task Design



Note. Participants remembered two items on every trial. Memory performance was manipulated in two ways: First, by presenting the items with many colored dots for a short time (high-noise trials; left) or a few colored dots for a long time (low-noise trials; right); more perceptual noise at encoding increases an item's confusability with other colors. Second, we presented a postcue (i.e., retro-cue) during the delay period that was either neutral (distributed-attention condition; right) or directed to one of the items with 100% validity (directed-attention condition; left). This attention manipulation determines which item is in the "focus of attention". These manipulations allowed us to modify representational fidelity and attentional focus relatively independently. Finally, each trial continued with a search task, followed by the memory report task. See the online article for the color version of this figure.

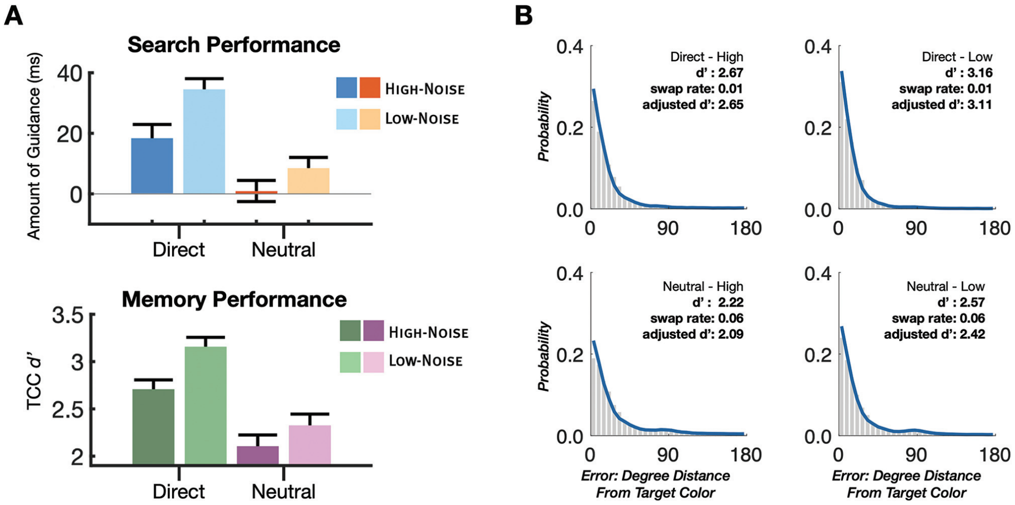
Model Fitting

Note that in theory, adding noise will change the similarity function used by the TCC model to estimate memory performance (d'), because we are changing the stimuli themselves, and thus changing the similarity structure of the stimulus space (Schurgin et al., 2020). In particular, if the noise makes colors more perceptually confusable, the central part of the error distribution might be wider with noise than no noise (as observed by Zhang & Luck, 2008). However, insofar as we are adding small amounts of color noise relative to the size of the stimulus, we assume these effects are small and continue to use the same color similarity function in the current data. This results in a good fit to the data (see Figure 7), suggesting the difference in similarity function is likely small with this level of noise. It may be that adding uniform color noise can be conceived of as simply increasing the familiarity of all of the colors present in the noise, and so with uniformly distributed noise, this is approximately equivalent to decreasing d'.

Results and Discussion

As predicted, manipulating attentional focus and perceptual noise at encoding significantly influenced memory performance. Memory performance (d') was submitted to a repeated measures ANOVA. The main effects of cue condition, $F(1, 49) = 182.81$, $p < .001$, noise condition, $F(1, 49) = 131.98$, $p < .001$, and search type (whether a tested memory item appeared in the search display; $F(1, 49) = 56.59$, $p < .001$) were significant. The interaction between the attention cue and noise conditions was also significant, $F(1, 197) = 17.16$, $p < .001$. Further analysis showed that average memory performance on trials with a directional cue (TCC $d' = 2.91$, swap rate = .01, adjusted $d' = 2.89$, circular $SD = 25.97°$) was superior to those with a neutral cue (TCC $d' = 2.39$, swap rate = .058, adjusted $d' = 2.26$, circular $SD = 34.05°$; $t(49) = 11.46$, $p < .001$; Figure 7B), indicating that a retro-cue is effective in increasing average precision of remembered items. Additionally, when an item was cued, memory performance was modulated by perceptual noise at encoding such that memory performance was higher when a low-noise item was encoded (TCC $d' = 3.16$, swap rate = .013, adjusted $d' = 3.11$, circular $SD = 23.8°$) compared with a high-noise item (TCC $d' = 2.67$, swap rate = .006, adjusted $d' = 2.65$, circular $SD = 28.05°$) and the difference between them was significant, $t(49) = 9.89$, $p < .001$. Memory performance was also affected by perceptual noise on neutral cue trials such that low noise trials resulted in better performance

Figure 7
Experiment 2 Results



Note. (A) Top: Amount of Guidance (difference in RT between target and distractor match trials) for each of the four tested-item-present conditions (trials of which we have both memory and search performance). When an item was placed within the focus of attention by a directional retro-cue, it biased attention in a graded fashion; more precise memory representations (low noise) resulted in a larger search effect compared with less precise representations (high noise). When no direct, attentional cue was present (neutral cue), the guidance effect was small overall. Bottom: Memory performance was highest for directional-cue trials with low-noise, followed by directional-cue trials with high-noise, neutral cue trials with low-noise, and finally neutral cues with high-noise. These data show that both the attentional cue and the perceptual noise at encoding modulated memory performance, and that this was directly related to the amount of guidance observed; consistent with a representational fidelity account rather than a special focus of attention. (B) Corresponding error histograms and TCC model fits for all conditions. See the online article for the color version of this figure.

(TCC d' = 2.57, swap rate = .059, adjusted d' = 2.42, circular Sd = 35.41°) compared with high noise trials (TCC d' = 2.22, swap rate = .055, adjusted d' = 2.10, circular sd = 38.7°; $t(49)$ = 8.24, $p < .001$).

Next, we looked at search performance and found that attention was most strongly guided by items with the highest memory quality. We submitted the search effect from tested-item-present trials to a repeated-measures ANOVA with retro-cue (neutral, directional) and perceptual noise (high-noise, low-noise) conditions. This analysis showed significant main effects of postcue, $F(1, 49)$ = 13.13, $p < .001$, perceptual noise, $F(1, 49)$ = 4.8, p = .03, and no interaction, $F(1, 49)$ = .61, p = .44. We performed planned follow-up t-tests to better characterize the search effect for items within the focus of attention, that were both searched and probed at the end of the trial, and found that participants were faster on target-match trials compared with distractor-match trials regardless of whether high-noise items, $t(49)$ = 2.46, p = .017, d_z = 0.35, or low-noise items, $t(49)$ = 5.50, $p < .001$, d_z = 0.78, were maintained. When attention was distributed between memory items we found a small search effect that failed to reach significance when a high noise item, $t(49)$ = 1.51, p = .14, d_z = 0.02, BF_{01} = 2.25, and a marginally significant effect when a low-noise item was maintained in working memory, $t(49)$ = 1.89, p = .06, d_z = 0.26, BF_{01} = 1.25.

Next, we looked at the difference in search effect across the directed attention conditions and found a reliable difference between high- and low-noise trials, $t(49)$ = 2.30, p = .02, d_z = 0.43. Given that these two trial types resulted in differential memory performance, with low-noise trials having better memory performance (TCC d' = 3.16) relative to high-noise trials (TCC d' = 2.67), this indicates that the quality of the memory representation—above and beyond attention alone—is important in determining the amount of guidance. This finding demonstrates that a template item guides attention in accordance with its representational fidelity. This is in contrast to what would be expected if "template status" per se was all that was critical to guidance.

Although the results from this experiment contradict the strongest versions of the attentional template account wherein an item's status alone determines which item guides attention (Fan et al., 2019; Hollingworth & Hwang, 2013; Zhang et al., 2011, 2018) it could be that the template status is a binary determinant of whether an item will guide at all, and how well an item is represented subsequently determines the strength of guidance. To more explicitly determine whether the attentional template is a necessary function of attentional guidance, we next investigate whether an item without a template status guides attention if it is represented with high fidelity in memory.

Experiment 3: Effects of Representational Fidelity on Items Outside the Focus of Attention

In a final experiment we tested whether an item that is not directly attended can guide attention, as long as its representational fidelity is sufficiently high; a prediction that is directly contradictory to attentional template accounts. To more directly compare the effects of representational fidelity (by adding noise at encoding) and attentional status (through a retro-cue) we manipulated both factors within the same trial. Attentional template accounts would predict no guidance for nontemplate items, whereas a representational fidelity account would predict that any well represented item exerts influence over attention.

Participants maintained one high-noise and one low-noise item in memory on the same trial and were subsequently retro-cued as to which item would most likely be tested on memory probe trials. Since it could be argued that memory performance is contaminated by the reappearance of one memory item in the search display, and this in turn may change participants' strategy during visual search and possibly interfere with the guidance effect, we now tested memory and search on separate trials (similar to our preliminary Experiments A1 and A2; see Appendix). This change eliminates any strategic attempt to improve memory by attending to the colors in the search display since participants will only perform one task per trial. We are particularly interested in whether a noncued and thus nontemplate item can guide attention when it is well represented, and whether, in general, d' tracks guidance, as we have seen in our previous experiments, and as predicted by our representational fidelity hypothesis.

Method

Participants

Owing to the global pandemic, in-lab data collection for this experiment stopped prematurely. We transitioned to an online study and, to make it more amenable to online testing, changed the length of the experiment, the sample size, and multiple aspects of the task itself (preliminary in-lab results mirror those found here). All participants were between 18 and 36 years old, reported normal or corrected-to-normal vision, and gave informed consent in accordance with the procedures approved by the Institutional Review Board at UC San Diego. 135 participants (72 women, mean age = 20.62 y) from UC San Diego took part in this online study in exchange for course credit. Exclusion criteria were identical to Experiment 1, except that we relaxed the search accuracy requirement to remove participants with worse than chance performance, and this resulted in the removal of 35 participants, giving a final sample size of 100 participants. With this sample size we can detect effects as small as $d_z = 0.28$, allowing for the possibility of the effect size being smaller owing to less reliable data than in the in-lab studies.

Stimuli

Participants performed the experiment on their home computers on a monitor that was at least 800x800 pixels to ensure the entire display was visible for the duration of the experiment. Similar to Experiment 2, two colored squares (90 × 90 pixels) were placed either side of fixation (300 pixels apart; centrally positioned 150 pixels to the left or right of fixation). Memory items were drawn from the same color space used in previous experiments and the color value for memory items were roughly 90° apart. Similar to Experiment 2, a circular cloud (50-pixel radius) of perceptual noise was superimposed over the memory colors. One memory item was occluded by 360 uniquely colored dots (4-pixel radius each; designated High-Noise item) and the other memory item was occluded by 60 uniquely colored dots (evenly spaced by 16° across the color wheel) and 300 dots that matched the color of the memory item (designated Low-Noise item; see Discussion). The search display was identical to previous experiments except where noted: search items were 150 pixels above or below fixation and contained either a straight line (distractor; 4 pixels wide and 55 pixels tall) or the target line which was identical to the distractor line except that it was titled by 30° either clockwise or counterclockwise. The search display briefly appeared for 200 ms before disappearing and showing only the fixation cross.

Procedure

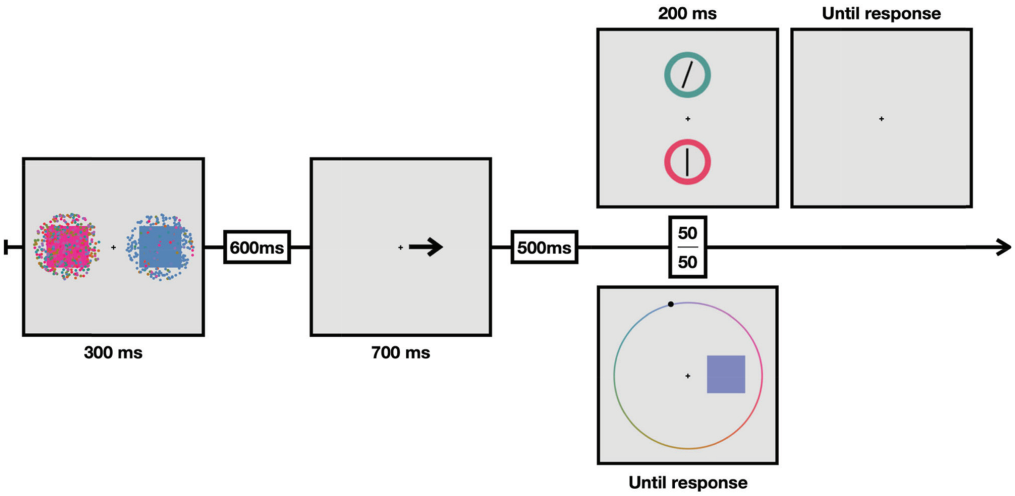
Participants performed a total of 320 trials that were evenly split between search trials and memory probe trials. On each trial two memory items were presented for 300 ms; one memory item was designated high-noise and the other item was designated low-noise. The location of each item varied randomly across trials. Participants maintained these items over a 600 ms delay prior to the presentation of a retro-cue (700 ms) which signified which item would be tested for memory with 80% validity (nonpredictive of search; 50% like to be present on search trials). The retro-cue was aimed at the high- or low-noise item evenly (160 trials each). After a final 500-ms delay the search or memory probe display was presented.

The search display was the same as in previous experiments (note: the search task was only displayed for 200 ms). Participants used the left or right arrow key on the keyboard to report the target orientation (counterclockwise and clockwise respectively). Immediately after a keypress, participants received feedback on their performance. On memory probe trials participants saw the memory probe display (see Figure 8) which informed them to report the cued or uncued memory item. Participants interacted with this task exactly as before; using the mouse to move around the continuous color wheel until the reported color matched their memory as closely as possible. Feedback was provided after a response was made.

Results and Discussion

In Experiment 3 we manipulated the representational fidelity of both items in working memory on the same trial and found (a) that the amount of attentional guidance follows the same pattern as the estimate of memory strength (d') of these items and (b) that uncued items—unattended items with no template status—guide attention as long as they are well represented (see Figure 9). For memory performance, we found significant main effects of perceptual noise, $F(1, 99) = 52.92$, $p < .0001$, and cue condition, $F(1, 99) = 43.7$, $p < .0001$, and no interaction, $F(1, 99) = .03$, $p = .86$. For search, we found significant main effects of cue validity, $F(1, 99) = 8.56$, $p < .001$, and noise manipulation, $F(1, 99) = 4.17$, $p = .04$, and no interaction, $F(1, 99) = .45$, $p = .5$.

Figure 8
Experiment 3 Task Design



Note. Participants remembered two items: one high-noise (left) and one low-noise (right) item across a short delay. A retro-cue signifies which item will be tested on (forced report) memory trials (80% valid on memory trials; not predictive on search trials). On half of trials participants performed a search task and on the other half of trials had their memory randomly tested (see Method for more details). See the online article for the color version of this figure.

In planned *t* tests we found a significant search effect for cued items (i.e., a template item); irrespective of whether the memory item was subjected to high-noise, $t(99) = 3.92$, $p < .001$, $d_z = 0.39$, or low-noise, $t(99) = 5.28$, $p < .001$, $d_z = 0.53$. Of particular interest, for uncued items (those without any template status) we found a significant search effect for low-noise items, $t(99) = 3.16$, $p = .002$, $d_z = 0.32$, but not for high noise items, $t(99) = .57$, $p = .56$; $d_z = 0.06$; $BF_{01} = 7.71$. These search results demonstrate that well represented items guide attention, that poorly represented items do not, and, critically, that an uncued item without a template status guides attention when it is well represented.

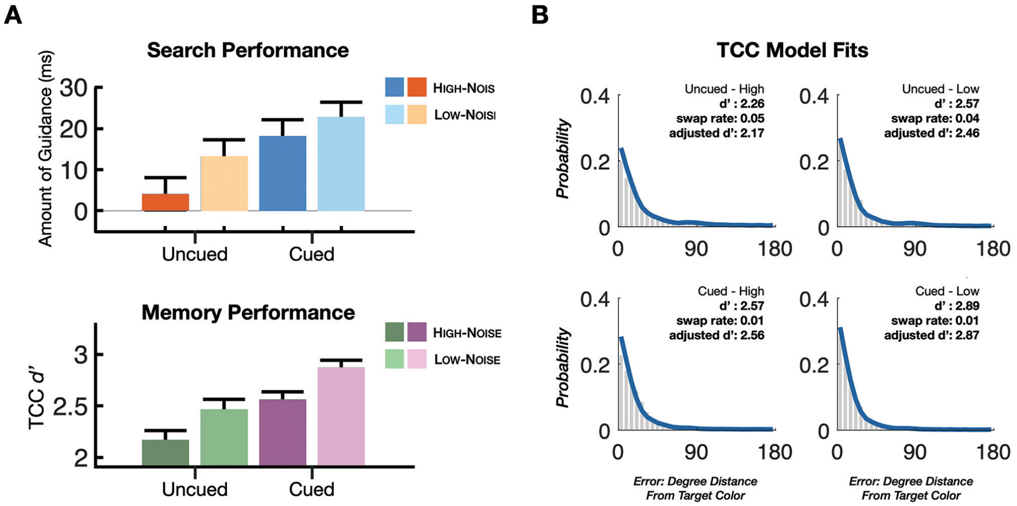
Although these results strongly argue against an attentional template account, it is plausible that when a poorly represented item is cued (high-noise-cued trials), participants do not focally attend to it and instead focus on the less noisy, uncued item. If true, we would expect a higher swap rate when the high-noise item was cued compared with when the low-noise item was cued since, if participants were internally attending to the wrong item, they would be more likely to report that item at test. However, we find no evidence to support this proposition as swap rate was low and very similar between high-noise and low-noise conditions (.041 vs .036, respectively, $t(99) = .43$, $p = .68$) and we found compelling evidence to support this null finding ($BF_{01} = 7.21$). These results lead us to conclude that participants maintained cued items as an attentional template (i.e., did not swap to better represented items) and that the search effect on the low-noise-uncued condition is from an item without a template status.

These results fit well with the results from Experiments 1 and 2 in suggesting that, on average, representational fidelity underlies the amount of observable guidance. They are also consistent with the idea that retro-cues are simply one way—a particularly effective way—among many others, to boost the representational fidelity of an item, which has downstream effects on guidance. Furthermore, these results are consistent with our representational fidelity account where any and all memory items guide attention insofar as they are well-represented. They are also, however, inconsistent with an attentional template account where only centrally focused item(s) can guide attention.

Explaining Guidance Effects Across All Experiments

Our proposed account makes a strong prediction, not just within experiments but also across experiments: memory strength is sufficient to explain attentional guidance, with no other factors needed (e.g., with all other factors exerting influence purely through their effect on representational fidelity). Our data are qualitatively consistent with this prediction: Experiment 1 suggests that within a trial, items that are best represented are most responsible for search effects, and Experiments 2 and 3 show that average memory performance predicts the average magnitude of the visual search effects, such that stronger memory representations were related to stronger guidance effects across trials. These conclusions include across-trial comparisons of memory performance where there was no possibility of reencoding the item during search and thus having the search display influence memory or vice versa (preliminary Experiments 1A and 2A and Experiment 3).

Figure 9
Experiment 3 Results



Note. (A) Top-Left: Search effect for each of the four noise and cue conditions. Even for items putatively "within the focus of attention" by a directional cue, more precise memory representations (low noise) resulted in a larger search effect compared with less precise representations (high noise). When an item was uncued (i.e., "outside of the focus of attention"; a "template status" had been given to the other item), guidance was similarly dependent on the representational fidelity of the item. That is, low-noise uncued items exerted robust guidance over attention, whereas high-noise (poorly represented) items did not. In short, the search effect followed memory performance regardless of whether an item was cued or not. Bottom-Left: Memory performance (adjusted *d*′) was highest for cued items that were encoded with less perceptual noise (low-noise) and worst for uncued items that were encoded with more perceptual noise (high-noise). (B) Corresponding error histograms and TCC model fits for all conditions. See the online article for the color version of this figure.

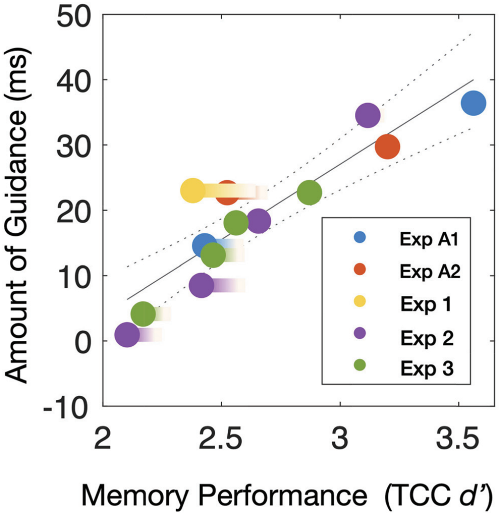
To quantitatively test the prediction that representational fidelity can explain the amount of attentional guidance, we correlated average memory performance with the average guidance effect across conditions from all of our experiments—including two supplementary experiments in which we varied set size (1 vs. 2, see Appendix). We included only forced report trials as they reflect the true underlying memory strength whereas free report trials overestimate the underlying fidelity of items as they are biased toward items that happen to have the highest representational fidelity on a particular trial. We submitted memory performance (TCC adjusted *d*′) and the amount of guidance (distractor-match minus target-match RT) to a simple linear regression model and found that memory performance was highly predictive of the guidance effect across experiments (Pearson's $r = .9$, $R^2 = .82$, $p < .0001$). For this analysis, we used adjusted *d*′ which assumes that individuals have no familiarity for the probed color when they report a color from the incorrect location; an assumption that results in conservative estimates of memory. Instead, it is likely that even when location information is lost, resulting in a swap, this deterioration does not lead to a total loss of information about the color (i.e., zero familiarity). Thus, we also plot the *d*′ values that presume memory strength is just as strong on swap trials to show the total possible range (the horizontal line connected to each point), and, as can be seen in Figure 10, this does not alter the nature of the relationship. This correlational analysis, with a clear increase to an item's ability to guide attention as that item is better represented, thus provides support for the hypothesis that the strength of memories is a critical factor in determining guidance effects.

In addition, as seen in Experiments 2 and 3, this analysis suggests that *very* strong memories are needed to guide search: The intercept of zero memory guidance is predicted to appear at approximately a *d*′ of 2, which is still a strong and extremely accurate memory. This may explain why natural variation between items in representational fidelity is sufficient to cause guidance to be largely driven by a single item, even at set size 2: even slightly deleterious noise is likely to remove the ability of an item to guide attention.

In summary, then, the model we propose can be instantiated as shown in Figure 11 This model has two parts: First, memories, even at the same set size, vary independently in representational fidelity (Fougnie et al., 2012; Panichello et al., 2018; Schurgin et al., 2020; Wilken & Ma, 2004). In the TCC model we use throughout, this is implemented via signal detection theory. Thus, even when both items at set size 2 are encoded with $d' = 2.0$, there is variation in the ultimate representational fidelity of these items, in part, because they accumulate independent noise (Figure 11, left). Second, our data strongly suggest that only items with high representational fidelity guide attention in a robust and reliable way. Figure 11 (middle) instantiates one particular version of this, where strong representational fidelity is required to guide attention.

Figure 10

Each Experiment Is Plotted as a Unique Color



Note. Multiple dots represent unique conditions. Dots correspond to memory performance: represented as the adjusted d' (d' for the probed memory item; a conservative measure which assumes participants have no familiarity for the probed item and decided to report the other one), because performance for the probed item is most relevant when asking about how memory relates to the search effect. Each dot has a corresponding colored line which represents the total possible range of memory strength: if participants maintained perfect memory representation for both items and simply reported the wrong location, d' would be at the far right of each line. In general, the amount of guidance increases with memory performance within and across experiments. See the online article for the color version of this figure.

Together, these two premises are consistent with the patterns of data we observe and that are observed in the literature. The particular numbers from Figure 11's instantiation are not necessarily fixed—they depend on various assumptions about what it means for only strong memories to guide, and how guidance might affect RT. However, the patterns remain the same regardless of these parameters. Such a model matches Figure 10 in terms of guidance as a function of average memory strength (d'). This model also predicts other effects we observe, like the heterogeneity between items even on the same trial. For example, at set size 2, with d' = 2.5, the model predicts that any given item has a 30% chance of causing guidance. However, since the noise for each item is independent, this implies that the chance both would guide attention is only 9% (30% × 30%), suggesting most trials will have only one item guide attention in a meaningful way.

General Discussion

Recent work has shown that attention can be biased toward items that match the contents of working memory. Using hybrid visual working memory and visual search paradigms, several studies have shown robust guidance when a single working memory item is maintained (Olivers et al., 2006; Soto et al., 2005, 2008; Soto & Humphreys, 2007, 2008) though the results are somewhat mixed when multiple working memory items are held in mind (Beck & Vickery, 2019; Chen & Du, 2017; Fan et al., 2019; Frătescu et al., 2019; Hollingworth & Beck, 2016; Hollingworth & Hwang, 2013; Houtkamp & Roelfsema, 2006; van Moorselaar et al., 2014; Zhang et al., 2018). The discussion of whether multiple items can guide attention is often focused on the number of items that can achieve a privileged template status with little focus on the representational fidelity of the remembered items. Here, in contrast, we carefully assessed the memory strength of items, demonstrating a straightforward relationship between the average representational fidelity of memories and attentional guidance that exists independently of an item's template status. In particular, we found that both within-trial and across-trial variation in representational fidelity predicted attentional guidance and did so without needing any other predictors (like a special privileged state). These findings suggest that the degree to which an item accurately represents the originally encoded item (i.e., its representational fidelity) determines whether—and how effectively—an item guides attention.

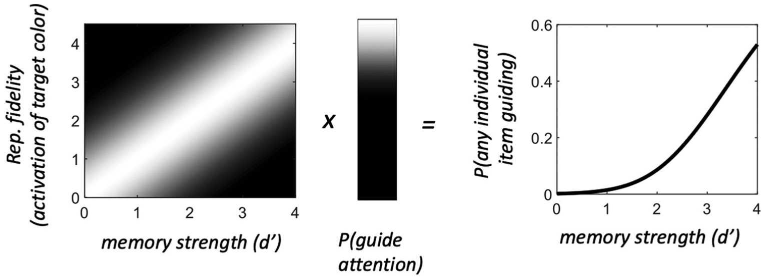
In particular, we show that when two items are maintained in visual working memory, one of the items tends to have greater representational fidelity than the other item, suggesting natural and inherent variation between memories; that the more accurate representation primarily drives the observed guidance effect while a poorly represented item exerts little to no influence over attention (Experiment 1); that the observed variation in memory reports between items is predicted by basic signal detection theories of memory (Simulation, Figure 5B); that, across-trial variation in the quality of memory representations predicts the size of search effects (Experiment 2); and, finally, that attention is guided by well represented items irrespective of achieving any purported template status (Experiment 3). Importantly, Experiment 3, along with the correlation across experiments (see Figure 10), suggests that multiple working memory items are each capable of guiding attention, as long as that item is maintained with sufficiently high representational fidelity (although this may be a rare occurrence in typical paradigms; see Figure 11). Although these latter results support a representational fidelity account, they are also in stark contrast with fundamental assumptions of the attentional template account.

The proposed representational fidelity framework speaks to two important issues in the literature of memory driven attentional guidance as well as to working memory and attention literature more broadly. First, to the question of whether one, or many working memory representations guide attention. Our data indicate that only an extremely strong and high-fidelity memory representation can guide visual search effectively, something that rarely occurs for more than one item at a time (e.g., Figure 10). To be clear, we do not suggest that multiple items could never guide attention simultaneously (as evidenced by Experiment 3), instead, the data simply suggests that all of the simultaneously maintained memories are unlikely to be sufficiently well-represented to each exert strong guidance over attention.

Second, the present results elucidate the mechanisms underlying attentional guidance and explain why attentional guidance is often driven by a single item. Importantly, and different from previous

Figure 11

Our Final Proposed Model Has Two Premises: Items Vary in Representational Fidelity, Even From the Same Display, Because of Independent Noise; and Only Strong Items Guide Attention



Note. Left: Even for items encoded with the same memory strength, the ultimate representational fidelity varies between items. In the TCC model we use to measure memory, this variation between items is a normal distribution with *SD* = 1, consistent with signal detection theory. Middle: Combined with the variation in precision between items is the fact that only strong items guide attention. The plot shows one possible instantiation of this, with the likelihood of attention being guided by an item based on the item's representational fidelity (here, Φ [*M* = 3.5, *SD* = 0.5]). Right: The result of this is that guidance occurs only when the items tend, on average, to be strong, matching Figure 10's real data.

accounts, our data suggest that natural variation in the representational fidelity between items is sufficient to explain the extent to which an item will guide attention on a particular trial, with no special focus of attention or similar state-based accounts of working memory being necessary (Experiments 2 and 3 and the Simulation). Under this account, retro-cues are simply one way to improve fidelity, but a similar boost in memory—and a corresponding boost in attentional guidance—can also be accomplished differently, as we show in Experiments 2 and 3.

The Importance of Representational Fidelity

The importance of strong memories and their inherent connection to an item's ability to guide attention has been acknowledged in the literature for many years. Unfortunately, however, this recognition has rarely translated into precise measurements of representational fidelity of individual items. For example, in an elegant set of experiments, Olivers et al. (2006) found no search effect when participants knew that the final memory test would be relatively easy (i.e., red vs green) but found a substantial search effect when participants knew that the memory test would be difficult (e.g., two subtle variations of red). While the goal of this manipulation was to assess the differences between verbally and visually maintained representations, the more difficult memory condition had the likely effect of producing higher representational fidelity for remembered items.

More importantly, much of the literature on attentional guidance does not use tasks that allow for direct measurement of memory strength for the relevant features at all. So, despite designing tasks that would encourage participants to maintain a highly precise representation, the memory probe itself cannot lead to an accurate estimate of representational fidelity. This is because tasks were used in which performance depends on memory for features that are not relevant for guidance (Chen & Du, 2017; King & Macnamara, 2020), or 2-AFC probes where the foil items are extremely similar to the target color (Hollingworth & Beck, 2016; van Moorselaar et al., 2014, Experiments 1–2 and 4). These manipulations make accurate assessments of memory strength impossible because they decrease estimated performance without changing the underlying memory signal, as shown by Schurgin et al. (2020). Furthermore, these manipulations are often combined with measures of memory that are not independent of response criterion, for example by averaging percent correct in a change detection task (e.g., Dube et al., 2019; Frătescu et al., 2019; Zhang et al., 2011)—a method which embraces a high-threshold model of responses, even though memory appears graded in nearly all studies, including in working memory studies (see Robinson et al., 2020).

In the case where studies use foils that are extremely similar to the target (e.g., Frătescu et al., 2019; Hollingworth & Beck, 2016; Olivers et al., 2006; van Moorselaar et al., 2014) their performance estimates effectively compress the performance scale—very strong memories are needed to get a *d′* above 0, and a *d′* of .5 in such a task might correspond to a *d′* of 3 or more in a task like ours or in a 2-AFC task with more distinct foils (Schurgin et al., 2020). This is to say that memory estimates from previous studies are not directly comparable for many reasons. Although Schurgin et al. (2020) demonstrate that 2-AFC tasks with maximally distinct foils effectively measure the same underlying memory strength as comparisons between more confusable colors or as continuous report, the nature of the seemingly low performance in some attentional guidance studies (e.g., 65% accuracy) leads many of these researchers to interpret memory as all-or-none (i.e., precise or not precise), potentially obscuring the relationship between memory strength and guidance.

Some attempts have been made to look at memory performance and guidance using more fine-grained measures. However, even findings which more accurately estimate memory have claimed that there is no relationship between representational fidelity and guidance when looking at individual trial errors (Hollingworth &

Hwang, 2013). However, correlational analyses that are performed on individual trials could never result in a meaningful or significant correlation since no models of memory support a direct linear relationship between error on a single trial and the underlying representational fidelity of that memory, especially not models where responses are inherently stochastic (e.g., Bays, 2015; Bays & Husain, 2008; Schurgin et al., 2020; van den Berg et al., 2012). For example, when sampling from the TCC model and assuming that d' was 100% perfectly predictive of guidance, the maximum correlation observed in such an analysis is $r = .09$ (because a given d' can result in any error, with only a slight change in their proportions). The null result observed by Hollingworth and Hwang (2013) and studies like it are thus not informative for the central issue of whether fidelity might underlie variations in guidance.

By contrast, in the current work we put a strong emphasis on accurately measuring memory quality for items and directly relate these measures to the guidance effects both within trials, across trials, and across experiments. This allows us to make clear predictions about which items guide attention and even allows us to quantify the representational fidelity that is needed for an item to guide while also determining how likely it is that more than one item exerts an effect during visual search (see Figure 11). Concurrent recent findings have also demonstrated how accurately estimating memory strength elucidates its importance to the magnitude of the guidance effect (Kerzel, 2019; Kerzel & Witzel, 2019). For example, Kerzel and Witzel (2019) find that a secondary working memory item does not guide attention and that this is not attributable to the lack of a template status but is simply attributable to that item being maintained with less representational fidelity than the guiding item. Similarly, Kerzel (2019) suggests that the number of guiding items is fewer than the overall capacity of working memory because the precision of guiding items must be extremely precise, not necessarily because a narrow (single item) attention template drives the effect. In the future, to further understand the role of representational fidelity in attentional guidance, it would be useful for those studying attentional guidance to use memory measures that precisely and accurately assess memory strength, and ideally measures that would allow for a comparison between studies. This could be achieved by using 2-AFC with maximally different foils at test, which would provide a measure of the upper bound of memory in these tasks (Brady & Störmer, 2021), or by reporting either TCC d' (Schurgin et al., 2020), or the circular standard deviation when using continuous report.

On the Number of Items That Guide Attention

Many studies have found guidance effects for one and two item working memory loads, including our supplementary experiments (see Appendix; Beck & Hollingworth, 2017; Beck et al., 2012; Hollingworth & Beck, 2016; Kerzel & Witzel, 2019; Olivers et al., 2006; Soto et al., 2005, 2008; Soto & Humphreys, 2007, 2008; Zhang et al., 2011). Such multiple-item guidance effects can always be explained in two ways: (a) both items equally guided attention, or (b) one working memory item is primarily responsible for driving the multiple item effect (Beck & Vickery, 2019; Downing & Dodds, 2004; Olivers et al., 2006, 2011; van Moorselaar et al., 2014; Zhou et al., 2020). Historically, dissociating between these two interpretations has been extremely difficult, as they mimic each other when averaging across trials. The representational fidelity account, described here, offers a new view on this question and has the potential to explain which cases would result in the guidance of attention by only a single item or multiple working memory items.

In many cases, the variation in fidelity across remembered items that we observe—and find to be extremely important for guidance—seems to undermine the strongest claims of two item guidance. For example, Hollingworth and Beck (2016) had participants maintain one or two working memory items while they searched for a single target among eight distractors. Using search displays with two distractors that could match one (match-1), both (match-2), or none (match-0) of the memory items, they showed that attention was guided on both match-1 and match-2 trials and found a greater effect when both items appeared in the search display. Although these results appear generally consistent with multiple items influencing attention, an alternative explanation is that because the best represented item was more likely to be present on the search display in the match-2 condition, a more consistent, and thus more robust, guidance effect was found. Specifically, on match-1 trials the best represented item would be expected to be present on 50% of trials, and on match-2 trials, the best represented item would be present on 100% of trials, which would generate a greater, and more reliable search effect in the match-2 condition on average, even if only a single item was guiding attention (a similar logic applies to a replication of this original study by Frătescu et al., 2019, and other studies with a similar design, for example, Fan et al., 2019; Zhou et al., 2020). Thus, in these types of tasks, the presence of random variation in fidelity between items, combined with only strong items guiding attention, potentially makes it difficult to know with confidence how much guidance is genuinely arising from the second item.

Of course, our account does not suggest that only one item is necessarily responsible for guiding attention; on some trials, when both items are represented extremely well, we predict that both items could exert at least some observable guidance. And, although this situation is unlikely to occur when participants are asked to remember colors that are randomly drawn from 360 unique colors, there are manipulations that could modulate memory strength to produce such an effect. A possible example of this principle is provided by a recent study by Chen and Du (2017) where they investigated whether two memory items could guide attention by combining two critical features from previous studies: a match-2 condition (Hollingworth & Beck, 2016) and a shape singleton search task (van Moorselaar et al., 2014). Across a convincing set of experiments, they provided data which suggested that multiple items can exert roughly equal guidance over attention. When participants remembered two items and were randomly presented with one of those items as a distractor (match-1), they found that attentional guidance was roughly equal when either memory item appeared. On trials where both items appeared as distractors (match-2), the attentional guidance effect (measured by their memory-driven capture index) was roughly double that of the match-1 condition and the guidance effect on match-2 trials was greater than when a single, cued memory item appeared in the search display. Thus, these findings appear to demonstrate that two items are capable of exerting roughly equal, additive guidance over attention. This finding is well explained by our representational fidelity account as it likely originates from the extremely well represented nature of the memory items: both items appeared to be represented with roughly equal precision (as measured by their 8-AFC task; see their Table 1), and memories were likely extremely strong because participants were shown two—of only four possible—unique colors[4] for long encoding time (1,000 ms) on every trial; likely supporting the creation of strong and less noisy memories.

Across our experiments, participants were shown a much larger stimulus set than is common in this literature (e.g., randomly selected colors from a continuous color wheel; 360 unique items). When two items were maintained, we found that memory strength for one of these items is substantially greater than the other and studies which have been taken to support a multiple-item template account have also shown this pattern for remembered items. For example, Zhang et al. (2018) used a paradigm similar to ours while also tracking eye movements: Participants remembered two colors (sampled from 180 unique values) while they performed a simple, two-item search task. Participants were cued as to which item would be tested first, and on half of the trials, participants reported their memory strength on a continuous color wheel instead of searching for a target among a single distractor. Although response time data from search trials was roughly equal for both memory items (cued and uncued), the authors measured first fixations as a more sensitive measure of a memory item's control over attention. While they conclude that multiple items guide attention, their data show that the proportion of first fixations tracked the reported memory strength of each item. That is, the cued item was fixated more often than the uncued item and the cued item was also maintained with greater fidelity. Their findings suggest that less well-represented items can interact with attention, but that they do so less efficiently and in direct relation to how well they are represented in memory. These results are similar to our data from Experiment 3 and, therefore, are in line with a representational fidelity account, which postulates that in principle any and all items can guide attention, but that the amount of guidance is determined by the underlying representational fidelity which can be modulated by attentional cues.

Representational Fidelity and the Focus of Attention

We show that differential memory performance between items (as indexed by different performance in forced vs. free report) arises in almost every situation (Exp. 1, and Supplemental Exp.), consistent with several other studies (Adam et al., 2017; Bays et al., 2009; Brady & Alvarez, 2015; Fougnie et al., 2012; Zhang & Luck, 2008). Why do memories vary in their representational fidelity, and how does this variation relate to attentional guidance? According to an attentional template account, one item is selected among other items, thereby getting a boost in memory performance while also gaining the ability to interact with attention. Why—and how—items are selected by attention and granted priority is often not specified, and it is an open question as to how this internal spotlight of attention acts upon memory representations especially with respect to when a representation is selected without top-down control or explicit instruction. The majority of previous work has used pre- and postcues to manipulate this focus of attention and has found that attended items guide attention more effectively, consistent with the focus-of-attention account (Olivers et al., 2011; van Moorselaar et al., 2014). However, in all of these studies the putative focus of attention and the quality of the memory representations were varied at the same time, because attended working memory items also resulted in better fidelity memories. Attentional cues changing fidelity could arise from a variety of sources, however, for example by devoting a greater proportion of resources to this item, or because it accumulates less noise as a result of being protected from interference (Bays & Taylor, 2018; Gazzaley & Nobre, 2012; Griffin & Nobre, 2003; Souza & Oberauer, 2016). Thus, it is possible that the focus of attention effects are simply effects of changes in memory quality.

Given that attentional priority typically covaries with improvements in memory performance, it is difficult to disambiguate between these two accounts. In Experiments 2 and 3, we designed a task that, for the first time, teased apart effects of attention and memory quality on attentional guidance. By adding perceptual noise at encoding, we manipulated the representational fidelity of each item. This memory manipulation was independent of any attentional cues, and by using a postcue after a short delay we manipulated priority, independently of the encoded precision. Our results showed that the postcue enhanced memory performance—as expected—and at the same time produced robust attentional guidance. Critically, however, we found that perceptual noise at encoding—which modulated memory quality but not attention—also influenced the guidance effect of attended items such that noisier items showed a smaller guidance effect, even when cued ("placed within the focus of attention," by such accounts; see Figure 10). Experiment 3, in particular, revealed that items that are not cued (i.e., outside the focus of attention and with no template status, by such accounts) also showed guidance effects as long as their memory was strong enough, as predicted by the representational fidelity account but wholly incompatible with an attentional template account.

Thus, we argue a straightforward and parsimonious explanation for differences in working memory guidance is that they arise from variations in the representational fidelity of items both within and across trials, and not because any item achieves a special status by being placed within a focus of attention. Rather, a framework in which representations vary in fidelity due mostly to independent noise but also modulated by display characteristics (e.g., Brady & Alvarez, 2015), cues, uneven resource allocation (Bays & Taylor, 2018), and more—but that does not require additional assumptions about differential states or templates to support item representations—can fully account for the present data, and a large set of data in the attentional guidance literature in general. This interpretation is also in line with signal detection models like the TCC memory model we use to fit our data, which predict natural variation in memory due to independent noise in the item representations (Schurgin et al., 2020). In fact, our simulations showed that the amount of variation in memory performance we find in our data is predicted by the natural stochasticity of signal detection theory, as implemented by TCC, with no need of any other explanations (like a focus of attention).

Because previous studies did not precisely measure differences in the representational fidelity between items and how that relates

[4] Note that the authors did not intend to use a small stimulus set and attempted to increase the reliance on working memory by including three textures that were superimposed over the colored disks. This resulted in a total of 12 items to-be-remembered. However, in such circumstances, color memory—the feature responsible for the guidance in this task—is generally almost always independent of other features (e.g., Fougnie et al., 2012), and so color memory was only tasked with remembering a total of four unique items.

to guidance effects, evidence for one item guiding attention was taken as evidence of a fundamentally distinct state of certain items in memory (and indeed this concept has been invoked to explain differential precision as well; see Oberauer & Lin, 2017). However, if items differ in how well they reflect the initially encoded item—owing to independently accumulated noise during the delay period—as they appear to do in nearly all working memory studies (Adam et al., 2017; Brady & Alvarez, 2015; Fougnie et al., 2012) and as we have shown here (Exp. A2 and Exp. 1), then it may be that poorly represented items do not bias attention simply because they are poorly represented. That is, as we show in Exp. A2, at set size 2, one item tends to be extremely well represented—just as well as the one item at set size 1—whereas the other item is much less precisely represented. If an item is poorly represented, it cannot, by nature, guide attention to the color that was previously encoded. Even when an item is cued, and putatively granted a privileged status, memory strength is a critical factor (Experiments 2 and 3) and the putative focus of attention cannot override the influence of representational fidelity (Experiments 2 and 3, see correlation Figure 10); with representational fidelity—and the effects of cues on such fidelity—instead seeming to be sufficient to explain guidance. Thus, our results do not provide evidence in favor of any attentional template account, even though they do show that in most cases one item primarily guides attention. Our data are largely consistent with a simpler view where attention is guided only to the extent that an item is well represented; with no added assumption of discreteness in memory states.

Conclusions

Selectively attending to relevant information in the environment is critical as we are subjected to more incoming sensory information than we could possibly process at once. Working memory allows us to maintain information no longer available to the senses for further processing, and it is imperative that these two systems interact successfully to navigate our environment. Here, we demonstrate that attention is biased toward objects that match the contents of working memory—even if task-irrelevant. Importantly, we show that working memory representations tend to guide attention only insofar as they are well represented, and that differences in representational fidelity between items is a natural process predicted by signal-detection theory. These findings have important implications for our understanding of the fundamental structure and processes involved in working memory and attention. Our interpretation of these results is that memory representations bias attention to the extent that they are well represented; this interpretation succinctly captures much of the data in the memory driven attentional guidance literature and does so without needing to invoke distinct states or special classes for working memory items.

References

Adam, K. C. S., Vogel, E. K., & Awh, E. (2017). Clear evidence for item limits in visual working memory. *Cognitive Psychology*, *97*, 79–97. https://doi.org/10.1016/j.cogpsych.2017.07.001

Bae, G. Y., Olkkonen, M., Allred, S. R., & Flombaum, J. I. (2015). Why some colors appear more memorable than others: A model combining categories and particulars in color working memory. *Journal of Experimental Psychology: General*, *144*(4), 744.

Bays, P. M. (2015). Spikes not slots: noise in neural populations limits working memory. *Trends in Cognitive Sciences*, *19*(8), 431–438.

Bays, P. M., & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science*, *321*(5890), 851–854. https://doi.org/10.1126/science.1158023

Bays, P. M., & Taylor, R. (2018). A neural model of retrospective attention in visual working memory. *Cognitive Psychology*, *100*(December 2017), 43–52. https://doi.org/10.1016/j.cogpsych.2017.12.001

Bays, P. M., Catalao, R. F. G., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision*, *9*(10), 1–11. https://doi.org/10.1167/9.10.7

Beck, V. M., & Hollingworth, A. (2017). Competition in saccade target selection reveals attentional guidance by simultaneously active working memory representations. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(2), 225–230. https://doi.org/10.1037/xhp0000306

Beck, V. M., Hollingworth, A., & Luck, S. J. (2012). Simultaneous control of attention by multiple working memory representations. *Psychological Science*, *23*(8), 887–898. https://doi.org/10.1177/0956797612439068

Beck, V. M., & Vickery, T. J. (2019). Multiple states in visual working memory: Evidence from oculomotor capture by memory-matching distractors. *Psychonomic Bulletin & Review*, *26*(4), 1340–1346.

Brady, T. F., & Alvarez, G. A. (2015). Contextual effects in visual working memory reveal hierarchically structured memory representations. *Journal of Vision*, *15*(15), 6. https://doi.org/10.1167/15.15.6

Brady, T. F., & Störmer, V. S. (2021). The role of meaning in visual working memory: Real-world objects, but not simple features, benefit from deeper processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Advance online publication. https://doi.org/10.1037/xlm0001014

Brainard, D. H., & Vision, S. (1997). The psychophysics toolbox. *Spatial Vision*, *10*(4), 433–436.

Chen, Y., & Du, F. (2017). Two visual working memory representations simultaneously control attention. *Scientific Reports*, *7*(1), 1–11.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–114.

Cowan, N. (2005). *Working memory capacity*. Psychology Press.

Downing, P. E., & Dodds, C. M. (2004). Competition in visual working memory for control of search. *Visual Cognition*, *11*(6), 689–703. https://doi.org/10.1080/13506280344000446

Dube, B., & Al-Aidroos, N. (2019). Distinct prioritization of visual working memory representations for search and for recall. *Attention, Perception & Psychophysics*, *81*(5), 1253–1261. https://doi.org/10.3758/s13414-018-01664-6

Dube, B., Lumsden, A., & Al-Aidroos, N. (2019). Probabilistic retro-cues do not determine state in visual working memory. *Psychonomic Bulletin & Review*, *26*(2), 641–646. https://doi.org/10.3758/s13423-018-1533-7

Fan, L., Sun, M., Xu, M., Li, Z., Diao, L., & Zhang, X. (2019). Multiple representations in visual working memory simultaneously guide attention: The type of memory-matching representation matters. *Acta Psychologica*, *192*, 126–137. https://doi.org/10.1016/j.actpsy.2018.11.005

Fougnie, D., Suchow, J. W., & Alvarez, G. A. (2012). Variability in the quality of visual working memory. *Nature Communications*, *3*, 1229. https://doi.org/10.1038/ncomms2237

Frătescu, M., Van Moorselaar, D., & Mathôt, S. (2019). Can you have multiple attentional templates? Large-scale replications of Van Moorselaar, Theeuwes, and Olivers (2014) and Hollingworth and Beck (2016). *Attention, Perception, & Psychophysics*, *81*(8), 2700–2709. https://doi.org/10.3758/s13414-019-01791-8

Gazzaley, A., & Nobre, A. C. (2012). Top-down modulation: Bridging selective attention and working memory. *Trends in Cognitive Sciences*, *16*(2), 129–135. https://doi.org/10.1016/j.tics.2011.11.014

Griffin, I. C., & Nobre, A. C. (2003). Orienting attention to locations in internal representations. *Journal of Cognitive Neuroscience*, *15*(8), 1176–1194. https://doi.org/10.1162/089892903322598139

Hollingworth, A., & Beck, V. M. (2016). Memory-based attention capture when multiple items are maintained in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(7), 911–917. https://doi.org/10.1037/xhp0000230

Hollingworth, A., & Hwang, S. (2013). The relationship between visual working memory and attention: Retention of precise colour information in the absence of effects on perceptual selection. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *368*(1628), 20130061. https://doi.org/10.1098/rstb.2013.0061

Hout, M. C., & Goldinger, S. D. (2015). Target templates: The precision of mental representations affects attentional guidance and decision-making in visual search. *Attention, Perception & Psychophysics*, *77*(1), 128–149. https://doi.org/10.3758/s13414-014-0764-6

Houtkamp, R., & Roelfsema, P. R. (2006). The effect of items in working memory on the deployment of attention and the eyes during visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(2), 423–442. https://doi.org/10.1037/0096-1523.32.2.423

Kerzel, D. (2019). The precision of attentional selection is far worse than the precision of the underlying memory representation. *Cognition*, *186*, 20–31. https://doi.org/10.1016/j.cognition.2019.02.001

Kerzel, D., & Witzel, C. (2019). The allocation of resources in visual working memory and multiple attentional templates. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(5), 645–658. https://doi.org/10.1037/xhp0000637

King, M. J., & Macnamara, B. N. (2020). Three visual working memory representations simultaneously control attention. *Scientific Reports*, *10*(1), 1–9.

Kiyonaga, A., & Egner, T. (2016). Center-surround inhibition in working memory. *Current Biology*, *26*(1), 64–68.

Klyszejko, Z., Rahmati, M., & Curtis, C. E. (2014). Attentional priority determines working memory precision. *Vision Research*, *105*, 70–76. https://doi.org/10.1016/j.visres.2014.09.002

Kumar, S., Soto, D., & Humphreys, G. W. (2009). Electrophysiological evidence for attentional guidance by the contents of working memory. *European Journal of Neuroscience*, *30*(2), 307–317.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279–281.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Erlbaum.

Morey, R. D. (2011). A Bayesian hierarchical model for the measurement of working memory capacity. Journal of Mathematical Psychology, *55*(1), 8–24.

Oberauer, K. (2002). Access to information in working memory: Exploring the focus of attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(3), 411–421. https://doi.org/10.1037/0278-7393.28.3.411

Oberauer, K., & Lin, H. Y. (2017). An interference model of visual working memory. *Psychological Review*, *124*(1), 21–59. https://doi.org/10.1037/rev0000044

Olivers, C. N. L., Meijer, F., & Theeuwes, J. (2006). Feature-based memory-driven attentional capture: Visual working memory content affects visual attention. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(5), 1243–1265. https://doi.org/10.1037/0096-1523.32.5.1243

Olivers, C. N. L., Peters, J., Houtkamp, R., & Roelfsema, P. R. (2011). Different states in visual working memory: When it guides attention and when it does not. *Trends in Cognitive Sciences*, *15*(7), 327–334. https://doi.org/10.1016/j.tics.2011.05.004

Ort, E., & Olivers, C. N. (2020). The capacity of multiple-target search. *Visual Cognition*, *28*(5–8), 330–355.

Panichello, M. F., DePasquale, B., Pillow, J. W., & Buschman, T. (2018). Error-correcting dynamics in visual working memory. BioRxiv. https://doi.org/10.1101/319103

Pelli, D. G., & Vision, S. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*, 437–442.

Pratte, M. S. (2019). Swap errors in spatial working memory are guesses. *Psychonomic Bulletin & Review*, *26*(3), 958–966. https://doi.org/10.3758/s13423-018-1524-8

Rajsic, J., Ouslis, N. E., Wilson, D. E., & Pratt, J. (2017). Looking sharp: Becoming a search template boosts precision and stability in visual working memory. *Attention, Perception & Psychophysics*, *79*(6), 1643–1651. https://doi.org/10.3758/s13414-017-1342-5

Robinson, M. M., Benjamin, A. S., & Irwin, D. E. (2020). Is there a K in capacity? Assessing the structure of visual short-term memory. *Cognitive Psychology*, *121*, 101305. https://doi.org/10.1016/j.cogpsych.2020.101305

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

Schneegans, S., Taylor, R., & Bays, P. M. (2020). Stochastic sampling provides a unifying account of visual working memory limits. *Proceedings of the National Academy of Sciences*, *117*(34), 20959–20968.

Schurgin, M. W., Wixted, J. T., & Brady, T. F. (2020). Psychophysical scaling reveals a unified theory of visual memory strength. *Nature Human Behaviour*, *4*(11), 1156–1172.

Soto, D., & Humphreys, G. W. (2007). Automatic guidance of visual attention from verbal working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(3), 730–737. https://doi.org/10.1037/0096-1523.33.3.730

Soto, D., & Humphreys, G. W. (2008). Stressing the mind: The effect of cognitive load and articulatory suppression on attentional guidance from working memory. *Perception & Psychophysics*, *70*(5), 924–934. https://doi.org/10.3758/pp.70.5.924

Soto, D., Heinke, D., Humphreys, G. W., & Blanco, M. J. (2005). Early, involuntary top-down guidance of attention from working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *31*(2), 248–261. https://doi.org/10.1037/0096-1523.31.2.248

Soto, D., Hodsoll, J., Rotshtein, P., & Humphreys, G. W. (2008). Automatic guidance of attention from working memory. *Trends in Cognitive Sciences*, *12*(9), 342–348. https://doi.org/10.1016/j.tics.2008.05.007

Soto, D., Llewelyn, D., & Silvanto, J. (2012). Distinct causal mechanisms of attentional guidance by working memory and repetition priming in early visual cortex. *Journal of Neuroscience*, *32*, 3447–3452.

Souza, A. S., & Oberauer, K. (2016). In search of the focus of attention in working memory: 13 years of the retro-cue effect. *Attention, Perception & Psychophysics*, *78*(7), 1839–1860. https://doi.org/10.3758/s13414-016-1108-5

van den Berg, R., Shin, H., Chou, W. C., George, R., & Ma, W. J. (2012). Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(22), 8780–8785. https://doi.org/10.1073/pnas.1117465109

van Moorselaar, D., Theeuwes, J., & Olivers, C. N. L. (2014). In competition for the attentional template: Can multiple items within visual working memory guide attention? *Journal of Experimental Psychology: Human Perception and Performance*, *40*(4), 1450–1464. https://doi.org/10.1037/a0036229

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision*, *4*(12), 11.

Wilson, B. M., Harris, C. R., & Wixted, J. T. (2020). Science is not a signal detection problem. *Proceedings of the National Academy of Sciences*, *117*(11), 5559–5567.

Woodman, G. F., & Luck, S. J. (2007). Do the contents of visual working memory automatically influence attentional selection during visual search? *Journal of Experimental Psychology: Human Perception and Performance*, *33*(2), 363–377. https://doi.org/10.1037/0096-1523.33.2.363

Zhang, B., Liu, S., Doro, M., & Galfano, G. (2018). Attentional guidance from multiple working memory representations: Evidence from eye movements. *Scientific Reports*, *8*(1), 1–9.

Zhang, B., Zhang, J. X., Huang, S., Kong, L., & Wang, S. (2011). Effects of load on the guidance of visual attention from working memory. *Vision Research*, *51*(23-24), 2356–2361. https://doi.org/10.1016/j.visres.2011.09.008

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*(7192), 233–235. https://doi.org/10.1038/nature06860

Zhou, C., Lorist, M. M., & Mathôt, S. (2020). Concurrent guidance of attention by multiple working memory items: Behavioral and computational evidence. *Attention, Perception & Psychophysics*, *82*, 2950–2962. https://doi.org/10.3758/s13414-020-02048-5

Appendix

Supplementary Experiments: Replications and Extensions

Before running the majority of the core experiments that support our fidelity account of attentional guidance, we ran several studies that sought to replicate basic findings that were necessary prerequisites of our account but that were not necessarily novel. Because these experiments do not provide new information but simply serve as the building blocks of our core experiments, we have put them in this appendix to reduce the burden on readers of the article. In particular, in Experiment A1, we verify that our paradigm finds guidance with both one item and two items in working memory, and greater guidance from one item than two. In Experiment A2, we verify that memories vary in representational fidelity in this paradigm, by including a free report memory condition where participants can report their strongest memory item, which we find results in better performance than forced report. In Experiment A3, we demonstrate that memory is a prerequisite of the guidance effect, and that priming per se does not drive the guidance effect in our experiments. Last, in Experiment A4, we demonstrate that the effect of attentional guidance is not exclusive to the two item search displays that we use throughout the main experiments.

Experiment A1: Visual Search Task With a One- or Two-Item Working Memory Load

Participants maintained either one or two colors in visual working memory and were asked to report a remembered color on 20% of trials using a continuous report color wheel. On the majority of trials (80%), instead of reporting the memory color, participants performed a visual search task in which the working memory color was irrelevant. Participants were instructed to report the direction of a tilted line (clockwise vs. counterclockwise from vertical). Based on previous research (e.g., Soto et al., 2005), we expected response times in the visual search task to be faster when the target was encircled by a color that matched the working memory color relative to when the distractor was encircled in the working memory color, reflecting the guidance of attention by a single visual working memory item. The main question was whether we would also find a guidance effect when participants maintained two colors in working memory.

Method

Participants

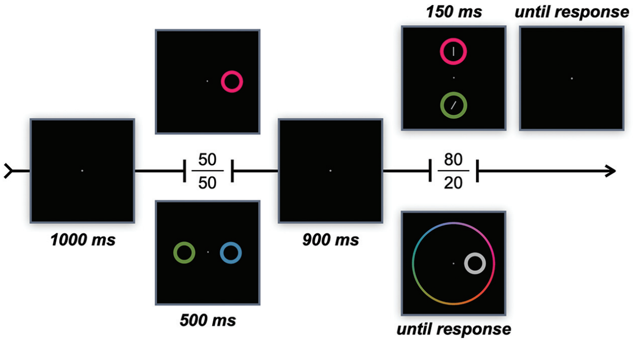
All participants were between 18 and 29 years old, reported normal or corrected-to-normal vision, and gave informed consent in accordance with the procedures approved by the Institutional Review Board at UC San Diego. Eighteen undergraduates (13 women, mean age = 21.22 y) from UC San Diego took part in this study in exchange for course credit. Data from two participants were excluded for poor visual search performance (< 50% accuracy). This sample size was determined to detect effects at least as small as $d_z = 0.7$ (with a power of .8 at $\alpha = .05$). Additionally, one participant reported the incorrect memory item (i.e., the color of the item at the opposite location, i.e., "location swap") on more than 40% of all trials (more than 3 standard deviations from the group average). As such a high swap rate suggests that the participant failed to perform the task as instructed—reporting colors independently of the probed location—we excluded data from this participant as well, leaving 15 participants in the final sample. The results and interpretation hold with or without the removed subjects.

Stimuli

The experiment consisted of color stimuli presented on a black background. Stimuli were generated and presented using MATLAB and the Psychophysics Toolbox (Brainard & Vision, 1997; Pelli & Vision, 1997). Memory items were colored rings that were 3° visual angle in diameter, .3° thick, and were centrally placed 4° to the left or right of fixation. On every trial, the color of one memory item was randomly drawn from a uniformly spaced circle (radius 49°) cut out of the CIE *L*a*b* space, centered at (L = 54, a = 21.5, b = 11.5) and when two memory items were present, the second color was selected to be 90° away in color space from the first color. The search display consisted of a target line which was .3° thick, .4° long, tilted .06° to the left or right of vertical, and placed 4° above or below fixation and a single vertical distractor line that was placed at the opposite location (see Figure A1). The target and distractor lines were encircled in colored rings that matched the

(*Appendix continues*)

Figure A1
Experiment A1 Task Design



Note. Participants were asked to remember either one or two colors (50%/50%) on each trial over a short delay and then either performed a visual search task that required them to indicate the tilt direction of a target line (80% of trials, top; target tilt is exaggerated compared with the experiment), or to report the color of one of the remembered items using continuous report (20% of trials; bottom). In the visual search task, the memory color was not predictive of the target location or orientation. See the online article for the color version of this figure.

memory item properties except for their color. One of the colors matched one of the memory colors and the other color was chosen to be 180° away from it in color space (at set size 2, this was 90° away from the other memory item). On the memory test display, one of the memory items was shown in gray (identical features to memory items) surrounded by a continuous color wheel which was 15° in diameter, .3° thick, and was centrally placed about fixation.

Procedure

Participants performed a total of 800 trials which were evenly divided between set size one and two. On each trial, one or two memory items were presented for 500 ms and participants were instructed to remember their color(s) as precisely as possible for a potential memory report task. After a 900-ms delay, participants performed either the visual search or the memory task. On 640 trials (80%) the search display was presented for 150 ms and participants reported whether the target line was tilted clockwise or counterclockwise from vertical by clicking the right or left mouse buttons, respectively. On 320 search trials (50%), a memory-matched color encircled the target line (target-match) while a distractor color, 180° away from the memory-matched color in color space, encircled the distractor line. On the remaining 320 search trials, a memory-matched color encircled the distractor line (distractor-match) while the distractor color encircled the target line. Thus, the memory color(s) never predicted which color the target line would be encircled by and was thus not useful for the search task. The location of the target and distractor line (top vs. bottom on the search display) was counterbalanced across the experiment. Feedback to respond more quickly was provided when responses exceeded 1,200 ms. On the remaining 160 trials (20%) participants were presented with a memory test display that consisted of a continuous color wheel and a single gray test-item placed to the left or right of fixation. Here, participants were asked to use the mouse to find the color closest to the remembered color on the color wheel. The location of the test-item indicated which memory item should be reported (e.g., a test-item on the left probed the color of the memory item that was on the left at encoding), and which item was tested was counterbalanced across the experiment. Once the mouse was moved from the central fixation point the gray test-item changed color to match the color at the position of the mouse cursor. Once participants identified the color that matched the remembered color as precisely as possible on the color wheel, they locked their response by clicking the mouse button. Response error, defined as the difference in degrees between the provided response and the correct answer, was shown after every memory trial and participants were instructed to keep this error below 10°. Participants were instructed to prioritize speed without compromising accuracy for the search task and, for the memory task, were instructed to prioritize precision without compromising temporal efficiency. On set size two trials (i.e., two memory items were presented at encoding), one of the memory items was randomly selected to be either the memory-matched color on search trials or the tested item on memory trials. All trial types were randomly intermixed within each block (see Figure A1). At the end of each block (40 trials) participants were shown an average of their memory response error and visual search performance (RT and accuracy) for that block.

Analysis

The analysis for preliminary Experiments 1 and 2 are identical to those described in the main text.

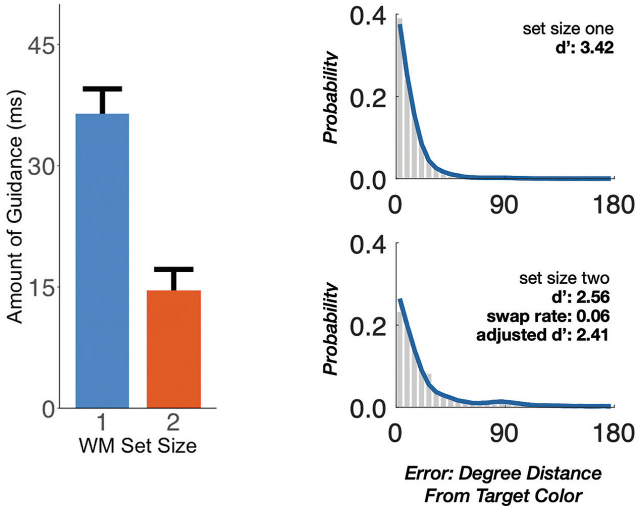
(*Appendix continues*)

Results and Discussion

The results of Experiment A1 suggest that attention is guided toward items that match those being actively maintained in working memory, even when working memory is loaded beyond a single item (see Figure A2 and Tables A1-A2). We submitted search-color condition (target-match, distractor-match) and set size (set size one, set size two) to a 2 × 2 repeated measures analysis of variance (ANOVA). This analysis revealed two main effects and an interaction, $F(1, 14) = 40.18$, $p < .001$, for the main effect of color condition; $F(1, 14) = 20.5$, $p < .001$, for the main effect of set size; $F(1, 14) = 12.15$, $p = .003$, for the interaction). Follow-up pairwise comparisons revealed that RTs were significantly faster on trials where the visual search target was encircled in a color that matched the working memory item (target-match) compared with when the memory-matched item encircled a distractor (distractor-match) for set size 1 trials, $t(14) = 6.94$, $p < .0001$; $d_z = 1.51$. A similar pattern of results, albeit with a smaller effect size, was identified when two working memory items were maintained, $t(14) = 2.61$, $p = .02$, $d_z = 0.71$. A speed–accuracy trade-off is unlikely since participants exhibited roughly equal accuracy on target-match compared with distractor-match search trials (set size 1: 94% and 92%, $t(14) = 2.07$, $p = .06$; set size 2: 94% and 93%, $t(14) = 1.63$, $p = .13$).

Despite probing memory on only 20% of trials, performance estimates were robust and overall quite good (set size 1: TCC $d' = 3.42$; circular $SD = 20.94°$; set size 2: $d' = 2.56$; swap rate = .06, adjusted $d' = 2.41$, $SD = 35.05°$). Performance was reliably lower at set size 2 relative to set size 1, $t(14) = 5.78$, $p < .001$ (see Figure A2). Participants appeared to "swap" and report the nontarget memory item nearly 6% of the time at set size 2, most likely because the intermediate task induced a loss of location information (mean swap rate for set size 2: .06; $t(14) = 3.65$, $p = .002$). As noted in the Method, we do not know the strength of memory for the target item on trials where participants misreported the nontarget item. Thus, we report d'

Figure A2
Experiment A1 Results



Note. Left: At both set size 1 and set size 2, search performance was faster when the target was encircled in a memory matched color (target match) compared with when it was encircled by a distractor color (distractor match; 180° away from a memory item), showing attentional guidance by working memory items. Right: Memory performance on the continuous report task. Memory strength was superior (higher d') when a single working memory item was maintained (top) compared with two working memory items (bottom). The gray bars reflect histograms of participant's errors, and the blue lines are the model fits. On about 6% of set size two trials, participants mistakenly reported the other (nonprobed) memory item, signified by the slightly elevated responses at 90° (location swaps). d' is the estimated memory strength for correct target reports only (ignoring swap trials), whereas adjusted d' reflects the memory strength when accounting for the likelihood that memory for the correct target was extremely weak when participants made location swaps. See the online article for the color version of this figure.

(*Appendix continues*)

Table A1
Average RT per Condition and Experiment

Experiment 1 (n = 30)	Forced Report: Item Present		Forced Report: Item Absent		Free Report: Item Present		Free Report: Item Absent	
	Target	Distractor	Target	Distractor	Target	Distractor	Target	Distractor
	578.21 (70.35)	600.58 (77.55)	580.13 (74.46)	603.77 (81.15)	582.05 (81.50)	602.11 (79.69)	589.01 (75.55)	591.47 (71.60)
Experiment 2 (n = 50)	**Direct Cue: High Noise**		**Direct Cue: Low Noise**		**Neutral Cue: High Noise**		**Neutral Cue: Low Noise**	
	Target	Distractor	Target	Distractor	Target	Distractor	Target	Distractor
	585.72 (65.49)	604.06 (65.94)	583.1 (66.65)	617.63 (65.01)	587.95 (63.02)	588.9 (61.58)	580.51 (56.79)	589.01 (64.18)
Experiment 3 (n = 100)	**Cued: High Noise**		**Cued: Low Noise**		**Uncued: High Noise**		**Uncued: Low Noise**	
	Target	Distractor	Target	Distractor	Target	Distractor	Target	Distractor
	561.12 (78.05)	579.19 (79.78)	566.59 (74.99)	589.28 (80.43)	578.42 (86.14)	582.55 (84.02)	573.3 (74.67)	586.43 (86.85)

Note. For each main experiment, and each condition, average RT is shown depending on whether a remembered item surrounded a target or a distractor (standard deviation in parentheses).

(memory performance on trials where the correct item was reported; i.e., assuming memory was exactly the same strength on swap trials) and adjusted d' (memory performance after adjusting downward to account for swaps, i.e., assuming target memories were nonexistent when participants made location swaps). These two estimates provide upper and lower bounds on the true memory strength of target items.

The visual search results of Experiment A1 replicate previous findings showing that working memory items can guide attention toward matching items in a visual search task. Furthermore, as expected, memory performance paralleled the observed search effects, with lower memory performance when two items were held in mind relative to a single item. Because we are averaging performance across trials, both the differences in search effects and the differences in memory performance for one versus two items could in principle be driven by overall less efficient processes when two items are held in mind relative to one, or alternatively, could be explained by one strong memory representation which also produced strong guidance on some trials, and one weaker memory representation, which presumably caused less guidance.

Experiment A2: Asymmetric Memory Strength for Multiple Working Memory Items

In this experiment we examined whether both memory items are represented equally well, or whether memory strength varies between items, such that one item tends to be represented better than the other item. In Experiment A1 we saw that both memory performance and attentional guidance are significantly decreased by holding in mind two items rather than one. This could occur because of averaging across trials and thus averaging over heterogeneity between items at set size 2, or because all items tend to be represented weaker in memory and thus result in smaller guidance effects.

We adapted the free report technique of Fougnie et al. (2012). Thus, on half of the set size two memory report trials participants were forced to report a randomly probed memory item (50% chance that either item would be tested, as in Exp. 1) and on the remaining trials, participants were free to choose one of the memory items to report (free report trials). These free report trials allow us to estimate the representational fidelity of a preferred item (preferred simply by nature of being selected) that is presumably the most precise item (Fougnie et al., 2012) and compare it with a randomly probed item.

Similar to Experiment A1, participants were shown either one or two memory items and performed either a visual search or a memory task on every trial, but on half of the set size two memory trials, participants were free to pick one of the memory items to report. On these trials placeholders for both memory items reappeared and participants clicked the location of the item they wished to report. If memory performance on free report trials resembles that of set size one, then we can conclude that variation between items in representational fidelity is large, and consistent with accounts where a single item drives

Table A2
Average RT per Condition and Experiment

Experiment	Set Size 1		Set Size 2	
	Target	Distractor	Target	Distractor
Experiment A1 (n = 15)	575.56 (66.05)	611.98 (62.19)	564.91 (66.07)	579.46 (71.40)
Experiment A2 (n = 18)	564.86 (99.62)	594.59 (94.34)	564.97 (90.42)	587.65 (91.39)
Experiment A3 (n = 65)			526.71 (168.88)	528.82 (161.42)
Experiment A4 (n = 65)			2,456.95 (806.33)	2,501.19 (813.57)

Note. These data are formatted as is in Table A1 above.

(*Appendix continues*)

the observed multiple item attentional guidance effect. However, if performance is more similar to set size 2: forced report, then it is less likely that a single item drives the effect and suggests that the reduced effect size for guidance at set size 2 is correlated with the reduced memory precision of the actively maintained working memory items—effectively supporting a multiple-item guidance account.

Method

The design, sample size, exclusion criteria, and analysis plan for this experiment were preregistered using AsPredicted (http://aspredicted.org/blind.php?x=nt3st3).

Participants

All participants were between the ages of 18 and 26, and the final sample included 18 undergraduates (10 women, mean age = 20.74y) from UC San Diego. Which, like Experiment A1, allowed us to detect effects as small as $d_z = 0.70$. Our preregistered exclusion criterion required high visual search performance, and in this case, this caused a large number of (13) additional participants to be removed and replaced for failing to achieve 80% accuracy in the visual search task. In addition, we removed and replaced 3 additional participants for having "swap rates" greater than 40% of trials, as in Experiment A1. We did not preregister this swap-based exclusion criterion. However, after analyzing the data, it became clear that these participants failed to follow instructions: such high swap rates indicate that these participants effectively performed "free report" on every trial, regardless of what location was cued, and thus do not provide useful data for distinguishing free versus forced report memory strength. Although overall this means a high number of participants were excluded (including 13 for the preregistered criterion and three for the swap-based criterion), including all of these participants in the final data set did not alter the results nor their interpretation.

Stimuli

All aspects of Experiment A2 were identical to Experiment A1 except for free report memory probe trials. On these trials, instead of one test item being cued, two gray circles indicating the possible test items appeared to the left and right of fixation (8° apart) prior to the presentation of the color wheel to allow participants the choice of which item to respond to (which they did by clicking the relevant location).

Procedure

The following aspects differed in Experiment A2 from Experiment A1. Memory probe trials were evenly split between three conditions: (a) set size one, (b) set size two: forced report, and (c) set size two: free report. All trial types were randomly intermixed throughout the experiment. On free report trials, at test, both items were presented, and participants were instructed to choose one memory item to report, either the left or the right. No further free report instructions were provided (i.e., participants were not incentivized or encouraged to select one item over another). Once a free report selection was made, the color that was encoded at that position was set as the correct response and response error was calculated in degrees as the difference between the correct and user selected response.

Model Fitting

We use the TCC model to fit both free report and forced report data. However, the actual response strategy in free report involves not just reporting the color with the strongest familiarity signal (as expected by TCC), but also comparing the two memory items and deciding which item has the stronger memory; something that is not instantiated in the TCC model that is fit to this data. Thus, the d' parameter for the free report fits will not reflect the intrinsic d' that each of the items are represented with but will instead be simply a description of the memory strength that would have been needed for a single item to match the results from the process of choosing the best represented item. The *Simulation* section following Experiment 3 in the main article addresses this in more detail.

Results and Discussion

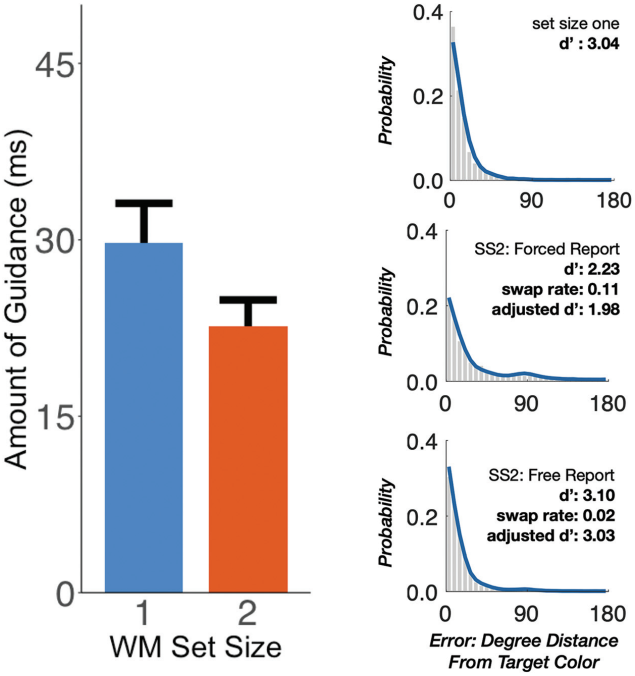
Replicating Experiment A1, we again found an attentional guidance effect both when one and two items were maintained in working memory. As in Experiment A1, color condition and set size were submitted to a 2 × 2 repeated-measures ANOVA. This analysis revealed a main effect of color condition, $F(1, 17) = 28.33$, $p < .0001$, but no main effect of set size, $F(1, 17) = .53$, $p = .48$, nor a reliable interaction, $F(1, 17) = 2.15$, $p = .17$. Participants were on average faster on target-match trials compared with distractor-match trials for both set size 1, $t(17) = 4.34$, $p < .001$, and set size 2, $t(17) = 4.85$, $p < .001$. Accuracy in this experiment was again quite good for target-match and distractor-match conditions and there was no evidence of a speed–accuracy trade-off (93% and 91%, $t(17) = 1.46$, $p = .16$; 93% and 92% $t(17) = 1.34$, $p = .20$; for set size 1 and 2, respectively).

With regard to memory performance, we found strong evidence in support of differential representational fidelity between the two items. In particular, when two items were held in working memory, one item was maintained as precisely as if only a single working memory item was remembered (see Figure A3). Thus, memory performance on free-report trials (TCC $d' = 3.10$, swap rate = .02, adjusted $d' = 3.03$, circular $SD = 27.42°$) is statistically indistinguishable from performance on set size 1 trials (TCC $d' = 3.04$, $sd = 27.93°$; $t(17) = .62$, $p = .55$; $BF_{01} = 3.40$; Rouder et al., 2009). Memory performance on random-probe trials (TCC $d' = 2.23$, $SD = 43.78°$) was comparable to performance from Exp. 1 and while we observed a slightly higher rate of location swap errors here (swap rate = .11) compared with Exp. 1 (swap rat*e* = .6) this difference was not significant, $t(31) = 1.92$, $p = .06$; $BF_{01} = 1.33$, suggesting that the inclusion of the Free Report manipulation did not result in a bias to preferentially attend to one compared with both memory items (consistent with validation of this report method from Fougnie et al., 2012).

These results show that, when two working memory items are actively maintained, one item has a stronger memory representation, resulting in considerably more precise color reports at test. This is consistent with accounts where

(*Appendix continues*)

Figure A3
Experiment A2 Results



Note. Left: Response times for target and distractor match trials in Experiment A2, separated by working memory set sizes. Replicating Experiment A1, we show robust search benefits for set size 1 and 2. Right: Errors and memory strength from the memory task, visualized with error histograms in gray and model fits in blue. As in Experiment A1, *d′* values reflect memory strength on correct-location report trials, and adjusted *d′* values reflect the assumption that participants had no memory for the target on swap trials, giving the range of possible memory strengths depending on assumptions about swaps. Performance on memory trials suggests that at set size 2, one item ends up with substantially greater representational fidelity compared with the other actively maintained item, as free report performance at set size 2 is as good as set size 1 performance and much better than forced report performance. See the online article for the color version of this figure.

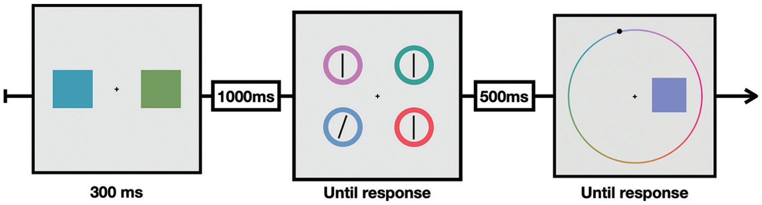
memory items are heterogeneous, either varying in precision intrinsically due to noise that accumulates independently over each item throughout the retention interval (e.g., Fougnie et al., 2012; Schurgin et al., 2020; Wilken & Ma, 2004), or as a result of a special focus of attention status (Oberauer, 2002; Oberauer & Lin, 2017). Furthermore, because one of the two items was maintained with set-size-one-like precision, these results are consistent even with strong versions of these accounts, where the multiple-item guidance effect is a mixture of two kinds of trials: a guidance effect, observed when the search trial contains the high-precision memory item; and a minimal, or nonexistent effect, observed when the search trial contains the secondary, low-precision memory item. Notably, however, unlike in Experiment A1, the guidance effect in this experiment was nearly the same size at set size 2 (d_z = 1.04) as at set size 1 (d_z = 1.21), which is inconsistent with this "mixture" account.

Experiment A3: Visual Search for Primed Colors

Previous work has shown that when the presented colors are no longer maintained in working memory that the attentional guidance effect disappears (Olivers et al., 2006). The guidance effect has also been absent when participants are simply primed, instead of needing to hold an item active in working memory (Kumar et al., 2009). However, it is feasible that memory is not a prerequisite for guidance and instead, the effects are driven by priming a sort of pop-out effect. To test this possibility, in Experiment A3,

(*Appendix continues*)

Figure A4
Task Design for Experiments A3 and A4



Note. In Experiment A4, participants were randomly probed on one of the remembered items. In Experiment A3 the task was identical except that memory was never probed. Instead, participants were simply instructed to attend to the colors in the encoding display. On every trial, four colored rings appeared and one of them contained a tilted line. Participants were asked to report the orientation of the line using the left and right arrow key on the keyboard. When memory was probed (Exp A4) a gray square appeared at the same location as the encoded item that was to-be-reported. Participants clicked the mouse when ready, and as they moved around the color ring, the probe changed colors. Participants locked in their response by clicking the mouse. See the online article for the color version of this figure.

participants were shown two colors and were told that they should simply attend to them before they disappeared. The effect observed in Experiment A1 was $d_z = 0.71$ when two items were maintained, and two items were searched. If the observed effect is genuine, we can reasonably expect it to shrink (see Wilson et al., 2020) especially in a task that is conducted online (compared with in-lab) and since searching a four-item search display is less efficient than a two-item display. Thus, if the within-subject effect is at least half of the original effect ($d_z = 0.35$) we would need at least 64 subjects to detect an effect at $\alpha = .05$ with a power of .8.

Method

Participants

All participants were between the ages of 18 and 35, and the final sample included 65 undergraduate volunteers (39 women, mean age = 21.37y) from UC San Diego who participated in this online experiment in exchange for course credit. Four subjects were removed for below chance accuracy.

Stimuli

The stimuli were identical to those used in Experiment 3 in all but the following ways. The four items in the search array were placed 150 pixels above and below fixation (and 150 pixels to either side like before). The three distractor colors were chosen to be at least ±70° (with ±15° additional jitter) away from the remembered color that appeared in the search display.

Procedure

Except where otherwise noted, this task was identical to Experiment 3 which was also conducted online. Participants performed a total of 384 trials that were split between target-match (25% of trials) and distractor-match trials (75%). On each trial two prime items were presented for 300 ms and one of these items was randomly selected to appear in the search display. After the encoding display disappeared, participants waited 1,000 ms before performing the four-item search task (which remained on the screen until a response was made; Figure A4).

Results and Discussion

In Experiment A3, we explored whether memory was a requirement of attentional guidance. Here, participants did not need to remember the presented items for a later memory test, instead they were instructed to simply attend to these items before they disappeared. Participants were not faster on target match trials compared with distractor match trials, $t(64) = .69$, $p = .49$, $d_z = 0.11$, $BF_{01} = 5.85$ (Figure A5) suggesting that priming is not sufficient per se to drive the guidance effect that we have observed in these experiments.

Experiment A4: Visual Search Task With More Search Items

In our other experiments, participants were required to search two items for a single target similar to previous work (Kiyonaga & Egner, 2015; Soto et al., 2012; Zhang et al., 2018). However, it could be that such simple displays are estimating some mechanism other than attentional guidance. In this experiment we had participants maintain two items in memory and search for a single target among three distractors. Similar to Experiment A3 we collected 65 subjects to observe an effect of $d_z = 0.35$ with a power of .8 (and an $\alpha = .05$).

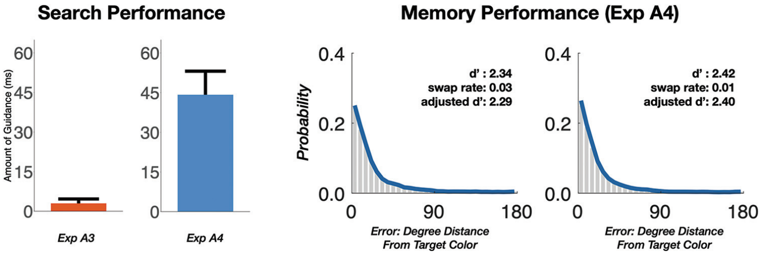
Method

Participants

All participants were between the ages of 18 and 28, and the final sample included 65 undergraduate volunteers (44 women, mean age = 20.55y) from UC San Diego who participated in this online experiment in exchange for course credit.

(*Appendix continues*)

Figure A5
Results for Experiments A3 and A4



Note. Left: Search performance, operationalized as amount of guidance (target—distractor match RT) for Experiment A3 (priming) and Experiment A4 (four item search). Replicating our previous experiments, we find a robust search effect when participants are required to search four items for a target and find no effect when participants are simply required to attend to the colors. Right: Memory performance for Experiment A4, errors and memory strength from the memory task at the end of each trial, visualized with error histograms in gray and model fits in blue. As before, model fits and d' represent average memory strength, and adjusted d′ values reflect he assumption that participants had no memory for the target when they incorrectly reported the nonprobed item, giving a range of possible memory strengths depending on assumptions about swaps. Memory performance is separated depending on whether the probed item appeared in the search display (50% of trials; Rightmost plot) or did not (Central plot). See the online article for the color version of this figure.

Eighteen subjects were removed from the final sample for failing to meet our previously used exclusion criteria. Seven subjects were removed for accuracy below chance while 11 subjects were removed for memory performance that was 2.5 standard deviations from the mean or who reported the other, nonprobed memory item more than 40% of the time (i.e., those with a swap rate exceeding 40%).

Stimuli

The stimuli were identical to those used in Experiments A3.

Procedure

Except where otherwise noted, this task was identical to Experiment A3 which was also conducted online. After responding to the search task, and after another 500-ms delay, participants were randomly probed on one of the remembered items. This probe was evenly split between the item that had just been searched and the passive item which was maintained for the memory task alone.

Results and Discussion

In Experiment A4 we increased the set size of the search display from two to four items and found an attentional guidance effect when participants maintained two items in memory (Figure A5). Participants were significantly faster when the target was surround by a memory item compared with when it surrounded a distractor, $t(64) = 2.50$, $p = .015$, $d_z = 0.31$. Memory performance improved marginally when the probed item appeared in the previously seen search display (TCC $d' = 2.42$, swap rate = .01, *adjusted* $d' = 2.40$, $sd = 28.1°$) compared with when it had not (TCC $d' = 2.34$, swap rate = .03, adjusted $d' = 2.29$, $SD = 28.5°$) and this difference in TCC d' was marginally nonsignificant ($t(64) = 1.98$, $p = .05$, $d_z = 0.25$, $BF_{01} = 1.18$). These results suggest that the guidance effect found in our other experiments are not the product of some undefined mechanism and are instead due to memory's guidance over attention.

Received May 17, 2021
Revision received October 14, 2021
Accepted November 15, 2021 ■

Acknowledgements

Chapter 1, in full, is a reprint of the material as it appears in the Journal of Experimental Psychology: Human Perception and Performance, 2022, Williams, Jamal R.; Brady, Timothy F.; Störmer, Viola S. The dissertation author was the primary investigator and author of this paper.

Chapter 2 Cutting Through the Noise: Auditory Scenes and Their Effects on Visual Object Processing

Jamal R. Williams and Viola S. Störmer

As it appears in

Psychological Science

2024

Research Article

ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Psychological Science
1–11

DOI: 10.1177/09567976241237737
www.psychologicalscience.org/PS
Sage

Cutting Through the Noise: Auditory Scenes and Their Effects on Visual Object Processing

Jamal R. Williams[1] **and Viola S. Störmer**[1,2]
[1]Department of Psychology, University of California, San Diego, and [2]Department of Psychological and Brain Sciences, Dartmouth College

Abstract
Despite the intuitive feeling that our visual experience is coherent and comprehensive, the world is full of ambiguous and indeterminate information. Here we explore how the visual system might take advantage of ambient sounds to resolve this ambiguity. Young adults (*ns* = 20–30) were tasked with identifying an object slowly fading in through visual noise while a task-irrelevant sound played. We found that participants demanded more visual information when the auditory object was incongruent with the visual object compared to when it was not. Auditory scenes, which are only probabilistically related to specific objects, produced similar facilitation even for unheard objects (e.g., a bench). Notably, these effects traverse categorical and specific auditory and visual-processing domains as participants performed across-category and within-category visual tasks, underscoring cross-modal integration across multiple levels of perceptual processing. To summarize, our study reveals the importance of audiovisual interactions to support meaningful perceptual experiences in naturalistic settings.

Keywords
Preregistered

Received 9/20/23; Revision accepted 2/15/24

In the real world, sounds are inexorably linked to the objects that generate them. Cats cannot bark and toads do not roar. In a world where visual features such as colors and orientations are inconsistent across viewpoints, lighting conditions, and time—where visual objects are often occluded and where many objects share similar visual features despite being fundamentally distinct—our perceptual system is required to constantly make inferences about the world (Alais & Burr, 2004; Bar, 2004; Körding et al., 2007; Oliva & Torralba, 2007). Context can help us to disambiguate indeterminate information: for example, the same shape projected on our retina might be interpreted as a hair dryer when viewed in a bathroom scene or as a drill when viewed on a workbench (Bar, 2004; Biederman et al., 1982). Similarly, visual scenes can facilitate the recognition of these objects quite dramatically (Davenport & Potter, 2004; Draschkow & Võ, 2017; Palmer, 1975). In the real world, however, sensory processing of visual scenes and visual objects is highly correlated, and a scene will often not provide independent or additional information (e.g., at dusk, all visual inputs are equally obscured). In this case, nonvisual information, such as sounds, can provide unambiguous and independent information about visual inputs, and potentially influence object recognition (Plass et al., 2017). But it is unclear whether naturalistic sounds facilitate object recognition or not, and if sounds do have this effect, what the mechanism might be. Further, it remains untested what kinds of auditory inputs—such as the sound of a specific object or the broader, ambient sounds of a scene—may influence visual object recognition. Therefore, in the present study, we investigate whether sounds of real-world objects and naturalistic scenes can facilitate how quickly relevant

Corresponding Author:
Jamal R. Williams, University of California, San Diego, Department of Psychology
Email: jrwilliams@ucsd.edu

visual information is extracted from noisy and ambiguous input.

Previous work has shown that when contextually relevant semantic information is provided in the form of written or spoken words, visual object processing is improved. For example, when written labels precede the faint image of a target object, they can increase the detectability of that object (Stein & Peelen, 2015), and when these labels are read aloud, they too facilitate visual processing and object identification (Lupyan & Thompson-Schill, 2012). In these cases, semantic information is thought to set up clear, top-down expectations about an upcoming object and then facilitate the processing of information that matches these expectations. However, the natural world rarely provides us with informative labels, so it is critical to understand how visual processing is influenced by nonvisual inputs that occur naturally in the environment—sounds that carry relevant and rich perceptual and semantic information and that thus represent reliable information that could support visual object recognition. To date, some studies investigating whether naturalistic sounds affect visual processing have found evidence consistent with this (Chen & Spence, 2010, 2011; Liu et al., 2012), though in some cases participants were first trained on sound-object pairs, which could possibly induce practice effects, as novel audiovisual relationships are learned and leveraged rapidly (e.g., Spence & Driver, 1994). Other work diverged from this, reporting that real-world sounds were ineffective or less effective than verbal labels in aiding visual object processing (Edmiston & Lupyan, 2015; Lupyan & Thompson-Schill, 2012). Recent functional magnetic resonance imaging (fMRI) results, however, point to a particularly interesting effect of sounds on visual processing: listening to naturalistic sound stimuli can activate the visual cortex in a content-specific way so that the activation pattern produced by sounds resembles the pattern produced by simply imagining the visual stimuli alone (Vetter et al., 2014). This suggests that abstract auditory information is fed to visual areas and modulates visual-cortical activity, perhaps in preparation of probable visual input. Additionally, the decoding of visual object identity is improved when visual objects are paired with task-irrelevant, congruent sounds (Brandman et al., 2020; de Haas et al., 2013), further suggesting that content-specific activity in visual cortex is modulated by sound. Although these findings suggest that auditory information modulates neural activity in visual regions, it is unclear how these findings might translate to behavior. This leaves two unanswered questions: How might naturalistic sounds influence the recognition of visual objects? And, in particular, can both auditory objects and auditory scenes affect visual object processing?

Statement of Relevance

Our perceptual system excels at navigating the complexities of the world by integrating information from different senses. Perhaps because the sound of an airport terminal often coincides with objects like luggage, the system can rely on this regularity to predict and facilitate the processing of such objects. In our study, participants viewed noisy visual objects while listening to naturalistic sounds. When sounds were related to the visual target, they facilitated the ability to extract relevant visual information, thereby accelerating object recognition. This was true for specific object sounds (a dog's bark) but also occurred for ambient auditory scene sounds (an airport terminal), indicating wide-ranging effects of audition on vision. Crucially, sounds aided categorical visual recognition (a dog from a bird) but also aided fine-grained visual discrimination (e.g., a malamute from a husky). Overall, our results demonstrate that sounds enhance vision across various levels of processing and stress the importance of cross-modal influences on perception.

We investigated these questions in a series of experiments in which we asked participants to perform an object-discrimination task while hearing naturalistic sounds (Williams et al., 2022). If sounds induce activity patterns in the visual cortex in a meaningful way, so that these activity patterns match, to some degree, the patterns that would have been induced by a visual stimulus, then relevant visual features might be given a processing head start, facilitating the extraction of those visual features that match the auditorily induced expectations. We are particularly interested in testing how broad the effects of naturalistic sounds on vision are. For example, do only sounds that are directly linked to a specific object support object recognition (e.g., the barking of a dog facilitates recognizing a dog)? Or can broader auditory information, such as the ambient sound of an environment (i.e., an auditory scene) influence how visual objects are processed? (E.g., does hearing the ambient sound of a park facilitate recognition of a park bench or a play structure?) Our task design allowed us to ask whether any influence of sounds on vision occurs at a rather broad, categorical level—such as helping to distinguish a bird from a train—or at the more detailed fine-grained level that is necessary for within-category distinctions, such as distinguishing a robin from a towhee. To anticipate our results, we found strong evidence that naturalistic

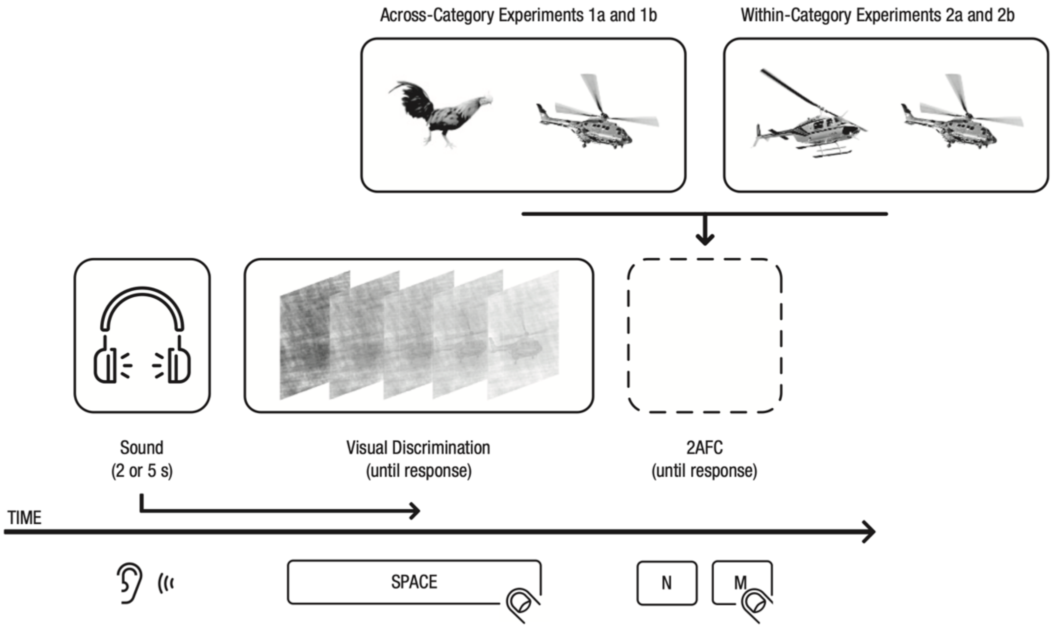


Fig. 1. Task design for all experiments. Participants first heard a real-world sound for 2 or 5 s (object and scene sounds, respectively). As the sound played, a target object slowly faded in through noise to become more and more visible. Participants were tasked with pressing the space bar to stop the visual-discrimination phase when they had sufficient visual information to perform the two-alternative forced-choice (2AFC) test at the end of each trial. In the 2AFC test, they chose which object they saw fading in through visual noise. In Experiments 1a and 1b the 2AFC test used across-category lures; in Experiments 2a and 2b the test used within-category lures. Response-time data during the visual-discrimination phase was the main dependent variable of interest.

sounds facilitate object recognition across all conditions, highlighting the important role of auditory context for visual perception.

Open Practices Statement

These studies have been preregistered, and the data, scripts, and stimuli needed to replicate or expand on these experiments and analyses are available online (https://osf.io/msqnv/).

Experiment 1: The Effects of Real-World Sounds on Visual Object Discrimination

Experiment 1 tested whether auditory objects or scenes influence how quickly a noisy visual object is recognized. Each trial began with a sound, and as the sound played, a visual noise patch continuously became less noisy to slowly reveal a target image. In Experiment 1a object sounds lasted for a total of 2 s, whereas scene sounds in Experiment 1b lasted 5 s. Participants pressed a button once they had acquired enough visual information to perform a two-alternative forced- choice (2AFC) test (Fig. 1). These features of the task ensured that any effect we might observe would not simply be due to a congruency bias, nor the reflection of speeded response preparation due to uncertainty (Heron et al., 2004).

Method

Participants. All participants in Experiment 1a gave informed consent in accordance with the procedures approved by the institutional review board at University of California, San Diego (IRB00000355), were between the ages of 18 and 26 years and reported having normal hearing and normal or corrected-to-normal vision. Initial piloting suggested an effect size of roughly .7, which would demand at least 18 participants with a power of .8 and an alpha of .05. We therefore collected 20 unique undergraduates (10 females; mean age = 20.8 years) who took part in this experiment in exchange for course credit.

This design, sample size, exclusion criteria, and analysis plan for Experiment 1b were preregistered on AsPredicted (https://aspredicted.org/VP2_1VN).

Presuming that the effect size may shift or shrink upon replication by roughly 75%, we estimated the need for roughly 30 participants for all future experiments. Thirty-nine unique undergraduates (18–38 years old, 29 females, mean age = 20.77 years) from our university took part in this experiment in exchange for course credit. Nine participants were removed on the basis of the preregistered exclusion criteria, leaving a final sample of 30 participants.

Stimuli. For Experiment 1a, real-world sounds were collected from online repositories and edited to be 2 s in length and have roughly equivalent amplitudes (within and across stimuli when played at roughly 70 db). Twenty-eight sounds were used in the main experiment with six additional sounds used exclusively for familiarizing participants with the task. This ensured that participants had no experience or practice with the stimuli used in the main experiment. Sounds in Experiment 1a included a wide range of objects, such as an ambulance siren, an acoustic guitar, an elephant's trumpet, and a teakettle (all stimuli available on the Open Science Framework). In Experiment 1b, we collected 49 auditory scenes that were edited to be 5 s in length. These sounds were then used in an online survey in which participants identified the scene (free-response format) while also providing information on what sorts of objects they expected to find in the scene. Nine sounds that were difficult to identify across participants were removed, and 40 sounds were used in the main experiment. The remaining nine sounds were used to familiarize participants with the task and were not used in the main experiment. We chose auditory scenes that were characteristic of ambient sounds present in common environments, such as a public park, a school graduation, a restaurant, and a football stadium. Scenes that were selected did not contain any discernible English and did not contain an object sound that would be used as a visual target object for the visual discrimination task. Next, we collected two images of real-world objects per sound, to create audiovisual categories (56 and 80 images for Experiments 1a and 1b, respectively; 500 × 500 pixels). For Experiment 1a, each image was chosen to match each sound's category (e.g., the sound of a car would be matched with an image of a car); for Experiment 1b, each image was chosen so that it would match an object that would likely be seen, but not heard, in the scene (e.g., the sound of a park was matched with a park bench). Visual objects were sampled from various online repositories.

Procedure. Visual stimuli were presented on a computer screen located approximately 60 cm in front of the observer, and auditory stimuli were presented in stereo through headphones. Participants performed a total of 56 or 80 trials (Experiment 1a and Experiment 1b respectively), and each sound was heard only twice—once as congruent and once as incongruent with the target image. Each trial began with the sound of a real object (Experiment 1a) or scene (Experiment 1b); Object sounds lasted for 2 s whereas scene sounds lasted for 5 s. For Experiment 1a, the target object began fading in immediately as the sound played. Sounds played for 2 s as the target object slowly faded into view. For Experiment 1b, sounds played for 500 ms before the image began to slowly fade into view, and the sounds continued for their entire duration or until a response was made (whichever came first). This would keep response times (RTs) and image clarity consistent across Experiments 1a and 1b and allow auditory information from the longer scenes to unfold. Participants were instructed to press the space bar on the keyboard as quickly as they could identify the object for the 2AFC task.

The initial visual-discrimination phase of each trial, in which an image faded in through noise, was designed the following way: Target images were stripped of their color and subjected to two types of noise: First, all images were layered (combined into a single image) and completely phase randomized, to create noise masks that would effectively obscure the target object. This full-noise mask contains image information from all images and thus obscures any apparent differences between the different images, serving as a particularly effective mask (cf. Störmer et al., 2019). Second, each image was phase scrambled individually, and on each trial this image was initially presented behind the full-noise mask. On each trial, the full-noise mask slowly became more transparent to reveal more of the underlying noisy target image. As this noise mask became more transparent, the phase randomization of the target object also slowly decreased in steps of 1% every 100 ms. In effect, the full-noise mask slowly faded away while the phase scrambling of the target image was reduced every 100 ms. Pilot experiments showed that participants could clearly identify each target object when the target image reached 70% clarity (7 s); therefore, the maximum length of a trial could be 7 s. Participants were encouraged to respond prior to this point, and trials in which participants waited the entire 7 s were removed from analysis. For this visual-discrimination phase, participants were told to "identify the object and press 'space' as quickly as possible." Before the experiment began, participants performed several practice trials in which they had to identify an object in noise, and each sound was completely irrelevant to the target object. They were given feedback on their accuracy on the 2AFC task and were given feedback on their response times in the visual-discrimination phase to encourage responses that were rapid without sacrificing accuracy.

In the main task, on half of all trials the sound was congruent with the target image, and on the remaining trials the sound and target image were incongruent. Thus, the sounds did not predict the images participants would see. Immediately after the visual discrimination phase, participants were presented with two images (200 × 200 pixels): the target and an across-category lure—for example, a car (the target) and a cash register (the nontarget lure). *Lures*, the nontarget images presented alongside the target, were selected at random across participants from the remaining (across-category) stimuli. Objects were presented on either side of a fixation point, and participants needed to press the "n" or "m" key to select whether the target from the visual-discrimination task was on the left or the right, respectively. After choosing, participants were presented with accuracy feedback on the 2AFC test before the next trial began. Last, in Experiment 1b, we included six sound-identification trials (randomly intermixed) in which participants had to identify the auditory scene that they had heard (open-response format). We used these probe trials to ensure that participants could (a) identify the auditory scenes and (b) were listening to the sounds.

Analysis. Our main analysis focused on comparing the mean RT during the visual-discrimination phase across congruency conditions for correct trials only (see preregistered analysis plan). We reasoned that these RTs reflect the process of accumulating visual evidence and expected that if sounds have an influence over this sensory-evidence-accumulation process, RTs should be faster for congruent relative to incongruent trials. Because participants were in control over how much of the visual object they saw, we did not expect any accuracy effects in the subsequent 2AFC task. We excluded trials with RTs that were 3 *SD*s away from an individual participant's mean. A participant's entire data set was removed if he or she (a) failed to achieve 65% correct on the 2AFC task, (b) if the average RT was faster or slower than 3 standard deviations from the group mean, or (c) if the participant was unable to correctly identify all six sound-recognition trials (for Experiment 1b only). Additionally, we performed two analyses for each experiment that were not preregistered: To verify that our observed effects were not driven by just a small subset of our stimuli, we performed a linear mixed-effects analysis that added stimuli as random effects to the fixed effect of congruency. We then compared this full model to a null model that excluded sound-target congruency as a factor. Individual stimuli were included in both models, so this allowed us to estimate whether congruency was the primary driver of our observed effects and not the variation present in the stimuli themselves. Second, we examined response times on the 2AFC test to validate our preregistered decision to analyze only correct trials. Because a small selection of participants demonstrated perfect accuracy, for these *t* tests we performed unequal variance with two sample tests (Welch, 1947). As a further analysis, we examined whether correctly identifying the sound of a scene had any correlation with the effect of congruency on response times. To do this between-subjects analysis, we took the correct identification rate of scenes from the initial survey and compared this with the congruency effect observed in Experiment 1b. To ascertain a correct identification, we took the free responses to the scene sounds and had three independent raters determine whether the response (a) exactly identified the scene (e.g., a carnival), (b) was thematically related (e.g., theme park, Disneyland), or (c) was unrelated and thus incorrectly identified (e.g., a stadium). We then compared the correct identification rate to the effect size that congruent sounds had over response times. Here, we found that the exact identification rate had no reliable correlation with the effect of congruency on response times, r = −.15, 95% confidence interval (CI) = [−0.50, 0.24], p = .44. It could be the case, however, that simply understanding the thematic relationship between the scene sounds and related objects is enough to produce a reliable correlation. To examine this, we considered both thematically related responses and exact responses as correct and again found no reliable correlation, r = −.26, 95% CI = [−0.58, 0.12], p = .18 (see Discussion).

Results of Experiment 1a

In Experiment 1a, we found that congruent object sounds sped RTs during the visual-discrimination phase compared to incongruent object sounds: Participants terminated the visual presentation earlier for matching sound-object pairs than for nonmatching sound-object pairs (4,029 vs. 4,245 ms, respectively), $t(19) = 3.24$, p = .004, Cohen's d = 0.72, 95% CI = [0.22, 1.21] (see Fig. 2). After we controlled for the variance introduced by the individual sound and target images in a linear mixed-effect model, the effect of congruency persisted, $\chi^2(1) = 7.25$, p = .007. This indicates that the congruency effect is not spuriously driven by just a few stimuli. To further ensure that the effect is not driven by a speed-accuracy trade-off, we compared accuracy across conditions and found no evidence that participants sacrificed accuracy for speed (d' = 2.44 vs. 2.26 for congruent vs. incongruent, respectively), $t(19) = 1.18$, p = .25, d = 0.26, 95% CI = [0.19, 0.71], BF_{01} = 2.34 (Fig. 3a).

One of our primary goals was to determine whether sounds facilitate robust and highly confident visual representations instead of testing whether sounds simply

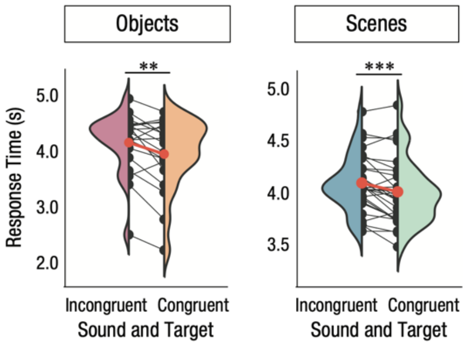


Fig. 2. Results for Experiment 1a (objects; left) and Experiment 1b (scenes; right). The average response time is shown in red. Individual participant response times for the visual-discrimination phase and the corresponding density plots demonstrate that participants were significantly faster when the sound was congruent with the visual target (orange and green distributions) compared to when the sound and target were incongruent (red and blue distributions).

modulate noisy or low-confidence visual inputs, which could induce decision or response biases. To that end, we preregistered the decision to analyze only correct trials, and when analyzing the differences between correct trials (547 ms) and incorrect trials (1,725 ms), we found that incorrect responses generally took longer in the 2AFC test, $t(14.337) = 4.12$, $p < .001$, $d = 1.04$, 95% CI = [0.4, 1.67] (see the Method section), irrespective of RT in the visual-discrimination phase (Fig. 3b). This supports our focus on correct trials, because response times are linked with confidence, accuracy, and the quality of a representation (Hellmann et al., 2023; Norman & Wickelgren, 1969; Ratcliff & Starns, 2013). This pattern was present across all experiments.

Results of Experiment 1b

Experiment 1b tested whether ambient sounds of scenes—the sound of a street or an airport terminal—can influence visual object recognition. Here, compared to the object-object pairing in Experiment 1a, where the relationship between auditory and visual stimuli is straightforward, the relationship between an auditory scene and any particular visual object is broader and indirect. In any given scene, there are typically dozens of sound generators, and there are many more objects that one might associate with that scene but which are not generally associated with a particular characteristic

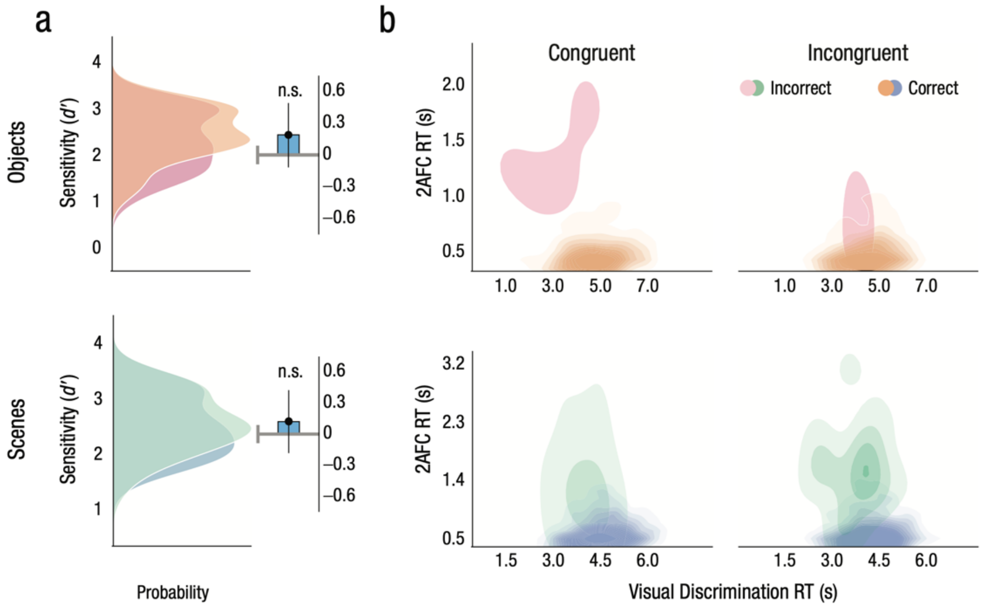


Fig. 3. Results for Experiment 1a (objects; top) and Experiment 1b (scenes; bottom). In (a), distributions of d' on the two-alternative forced-choice (2AFC) test were overlapping for congruent trials (orange and green; top and bottom) and incongruent trials (pink and blue; top and bottom). We found no significant difference in d' between the congruency conditions. Across both experiments, correct responses in the 2AFC test (b) were accompanied by faster 2AFC response times (RTs; orange and blue); incorrect responses were accompanied by slower and more variable 2AFC RTs (pink and green).

sound (e.g., for the sound of a park: a tree, a pond, a bench, etc.). Therefore, if scenes were to activate relevant visual features, this might result in the relatively broad activation of associated visual features. However, irrespective of this potential challenge to visual perception, in Experiment 1b we found that auditory scenes produced shorter RTs on congruent compared to incongruent trials (4,036 vs. 4,125 ms, respectively), $t(29) = 4.03$, $p = .0004$, $d = 0.74$, 95% CI = [0.33, 1.14] (Fig. 2), indicating that participants required less visual information to perform the 2AFC test on congruent compared to incongruent trials. Just as in Experiment 1a, this difference in speed did not result in significant differences in performance on the subsequent 2AFC test, $d' = 2.55$ vs. 2.48, $t(29) = 0.49$, $p = .63$, $d = 0.09$, 95% CI = [0.27, 0.45], $BF_{01} = 4.61$ (Fig. 3a), indicating that there was no speed-accuracy trade-off. We also conducted an item analysis to ensure that our observed effect was not spuriously driven by the variability of our stimulus set. As in Experiment 1a, we found that the effect of congruency persists even after accounting for this variability, $\chi^2(1) = 9.48$, $p = .002$. When examining RTs for correct trials (657 ms) and incorrect trials (2,079 ms) we found a similar pattern as in Experiment 1a, $t(27.72) = 6.25$, $p < .001$, $d = 1.19$, 95% CI = [0.70, 1.67] (see Fig. 3c, bottom).

Experiment 2: The Effects of Sound on Detailed Visual Object Discriminations

Thus far participants required only relatively coarse categorical information to perform the 2AFC test. Therefore, in Experiment 2, to test whether sounds support the extraction of fine and specific visual details about these objects, we implemented a within-category 2AFC test that required detailed visual information about the objects (Fig. 1).

Method

The design, sample size, exclusion criteria, and analysis plan for both Experiments 2a (https://aspredicted.org/VFV_44W) and 2b (https://aspredicted.org/H1P_PL7) were preregistered using AsPredicted.

Participants. As before, our final sample was preregistered to include 30 undergraduates who performed these tasks in exchange for course credit. For Experiment 2a, 32 unique participants were recruited (18–32 years, 12 females, mean age = 20.06 years), and data from 2 participants were removed; for Experiment 2b, a separate group of 34 participants was recruited (18–30 years old, 19 females, mean age = 21 years), and data from 4 participants were removed for failing to meet our inclusion criteria. Our preregistered analysis plan included running primarily *t* tests to compare mean RT for correct congruent and incongruent trials. However, to ensure that our results were not unduly influenced by the variance inherent to our stimulus set, after preregistering, we added a linear mixed-effects analysis to each experiment, as in Experiment 1. For our replication of Experiment 2a, we recruited a new sample of 20 participants from the undergraduate pool (no participants were removed for failing to meet our inclusion criteria).

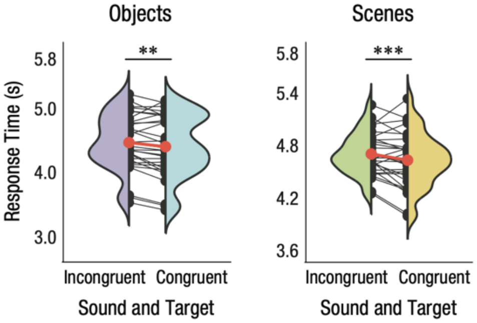


Fig. 4. Results for Experiment 2a (objects; left) and Experiment 2b (scenes; right). The average response time is shown in red. Individual participant response times for the visual-discrimination phase and the corresponding density plots demonstrate that participants were significantly faster when the sound was congruent with the visual target (blue and yellow distributions) compared to when the sound and target were incongruent (purple and green distributions).

Stimuli and procedures. All stimuli and procedures were identical to the previous experiments except that we added two extra audiovisual sets for Experiment 2b, and both images from an audiovisual set were used as target and lure on the 2AFC task, thus creating a harder, within-category discrimination than in Experiments 1a and 1b. That is, when participants saw a helicopter as a target object, the 2AFC would present the helicopter they saw during the visual-discrimination phase and a lure that was also a helicopter (see Fig. 1, top right).

Results of Experiment 2a

When participants heard object sounds and had to perform a more difficult, within-category 2AFC test, we found faster RTs for congruent compared to incongruent audiovisual trials for the visual-discrimination phase (4,454 vs. 4,524 ms, respectively), $t(29) = 2.78$, $p = .009$, $d = 0.51$, 95% CI = [0.12, 0.88] (see Fig. 4), which is similar to our findings in Experiment 1a. This speed difference did not have a significant effect on 2AFC performance, $d' = 1.89$ vs. 1.85, $t(29) = 0.32$, $p = .75$, $d = 0.06$, 95% CI = [0.30, 0.42], $BF_{01} = 4.90$ (Fig. 5a). After

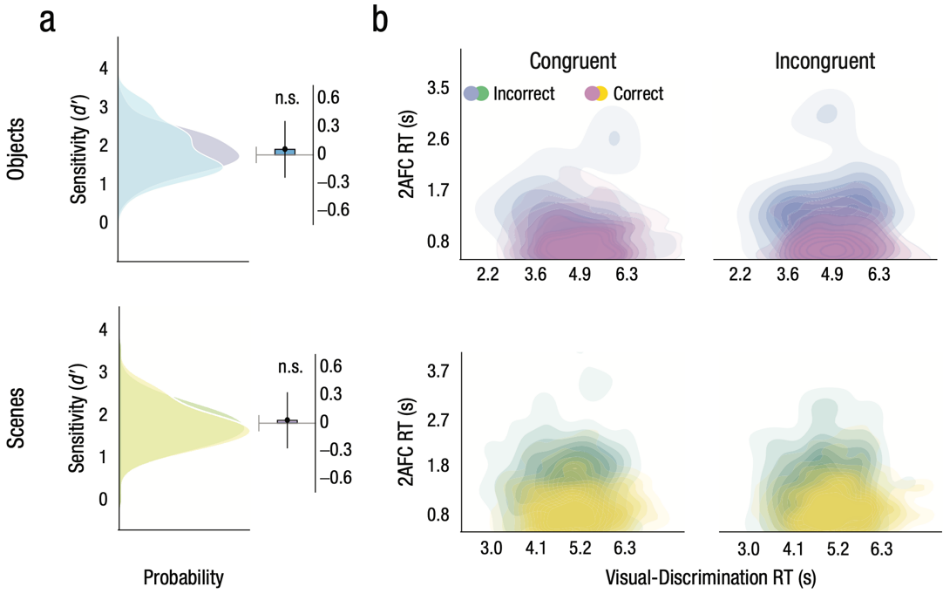


Fig. 5. Results for Experiment 2a (objects; top) and Experiment 2b (scenes; bottom). In (a), the distribution of d' on the two-alternative forced-choice (2AFC) test were overlapping for congruent trials (blue and yellow; top and bottom) and incongruent trials (purple and green; top and bottom). Performance was quite high, and we found no significant difference in d' between the congruency conditions. Across both experiments, correct responses in the 2AFC test (b) were accompanied by faster 2AFC response times (RTs; purple and yellow), whereas incorrect responses were accompanied by slower and more variable 2AFC RTs (blue and green).

further analysis of RT during the 2AFC, we found that participants were slower when they were incorrect (1,412 ms) compared to correct (981 ms), $t(38.64) = 3.95$, $p < .001$, $d = 1.08$, 95% CI = [0.61, 1.54]. Although we found faster RTs for congruent compared to incongruent trials on average (our preregistered analysis), we found that the variance in our stimulus set may have contributed to the significance of this effect. That is, once we accounted for stimulus variance in our model, the congruency condition explained a marginally nonsignificant portion of the variance, $\chi^2(1) = 3.53$, $p = .06$.

Although our results supported our preregistered analysis plan, we found in Experiment 2a that stimulus variability contributed to the observed effect of congruency. Because our paradigm was specifically designed to capture the potential effects on vision of sounds that occur naturally in the world—with only 60 trials and no training with the experimental stimuli—this may have been a power issue. Thus, to test this and validate our results from Experiment 2, we performed a replication in which we roughly doubled the number of trials. Here, the stimuli and procedure were identical except that we added more object sounds to the original set and heard each sound four times (twice as congruent and twice as incongruent), thus performing 128 trials overall, more than twice the amount compared to all other experiments. With these changes, we replicated the RT effect (i.e., with congruent trials being faster than incongruent trials; 3,721 vs. 3,805 ms), $t(19) = 2.38$, $p = .03$, $d = 0.53$, 95% CI = [0.06, 0.99]. We found no difference in d' ($d' = 2.23$ vs. 2.41), $t(19) = 1.07$, $p = .3$, $d = 0.24$, 95% CI = [−0.21, 0.68], $BF_{01} = 2.59$, and found that congruency was a significant factor once the variance in our stimulus set was accounted for, $\chi^2(1) = 9.48$, $p = .002$, thus replicating Experiment 2a.

Results of Experiment 2b

In Experiment 2b, we used auditory scenes and a within-category object-discrimination task, and we found that congruent sounds facilitated the extraction of visual details of objects: Participants were faster with congruent sounds than with incongruent sounds (4,674 vs. 4,762 ms, respectively), $t(29) = 3.59$, $p = .001$, $d = 0.66$, 95% CI = [0.26, 1.05] (Fig. 4). This difference in speed did not have a significant effect on accuracy, $d' = 1.69$ vs. 1.69, $t(29) = 0.04$, $p = .97$, $d = 0.01$, 95% CI = [−0.36, 0.35], $BF_{01} = 5.14$ (Fig. 5). As in Experiment 2a, responses were slower on the 2AFC task when participants were incorrect (1,762 ms) compared to when they were correct (1,196 ms), $t(41.11) = 4.72$, $p < .001$, $d = 1.29$, 95% CI = [0.79, 1.77]. Last, once we accounted for the variance in our stimulus set, congruency between the audiovisual pairs still accounted for a significant amount of the remaining variance, $\chi^2(1) = 12.33$, $p = .0004$. These results suggest that the broad and abstract information provided by auditory scenes can facilitate the accumulation of detailed visual information necessary to perform a within-category discrimination task.

General Discussion

The current study tested whether naturalistic sounds affected visual object processing by accelerating how quickly relevant visual features are extracted from noisy visual input. We used a visual-discrimination task and a perceptual test of object recognition to demonstrate that auditory information can hasten visual-feature extraction without negatively impacting accuracy. With this paradigm, we show that participants responded more quickly in the visual-discrimination phase when a nonpredictive sound was congruent with the target image compared to when it was incongruent, suggesting that participants demanded a greater amount of visual information to perform the 2AFC test for incongruent compared to congruent trials.

Whereas previous research examined the influence of auditory objects on visual object processing, reporting mixed results (Chen & Spence, 2010; Edmiston & Lupyan, 2015; Schneider et al., 2008), it was entirely untested whether auditory scene information would influence visual object recognition. Our results reveal that congruent auditory object and scene information facilitate visual object processing when compared to incongruent sounds, and they further demonstrate that these cross-modal effects occur for both categorical and detailed object recognition. This suggests that naturalistic sounds are not simply supporting broad visual categorization (i.e., is it a bird or not?) but are also helping to extract detailed visual-feature information as well (i.e., is it this bird or that one? Experiment 1b–2b). Taken together, these experiments reveal that visual perception is not only affected by the direct relationship between visual objects and the sounds that they make but that more abstract auditory information is leveraged as well to enhance the extraction of meaningful and detailed visual features.

However, this work cannot unambiguously determine whether congruent sounds facilitate, or incongruent sounds hinder, visual processing. Future work should fully explore whether the effect is driven by facilitation, hindrance, or a combination of both. Despite this, faster response times together with no cost in accuracy could be interpreted as a reflection of congruent sounds accelerating the extraction of meaningful visual information when compared to incongruent sounds. Because equivalent performance on the 2AFC task was observed when less visual information was available, this perhaps represents an improvement of visual sensitivity, as seen in previous work (Chen & Spence, 2011; Lupyan & Ward, 2013; Meyerhoff & Huff, 2016). Because the auditory objects and scenes contained reliable and unambiguous information, this may have generated relatively automatic predictions about plausible incoming visual information (Stein & Peelen, 2015; Winkler et al., 2009) which were then leveraged by the visual system (Oliva & Torralba, 2007). Although automaticity was not directly examined here, the sounds in our task were not predictive of what visual object a participant would see, and participants heard each sound only twice, making it unlikely that participants would set up a task-specific, top-down volitional expectation about what object they would see. This, in our view, suggests that these audiovisual interactions may occur relatively effortlessly. Although this interpretation is broadly consistent with recent neuroimaging studies showing that both auditory objects and scenes automatically drive the activation of congruent visual information neurally (Mulatti et al., 2014; Vetter et al., 2014, 2020), we cannot fully dismiss the potential of top-down, volitional contributions on the basis of the current results.

A top-down account should predict that identifying sounds is critical to set up expectations that are then explicitly used to search for expected input (e.g., Stein & Peelen, 2015). Although the direct relationship between auditory objects and their visual counterparts makes this mechanism feasible, auditory scenes have no direct relationship with any particular visual object, as objects are often probabilistically, not deterministically, related to their scenes (Greene, 2013).

This makes this mechanism less likely for scenes because they are harder to identify, and one would need to set this expectation for many objects and features. However, one way to explore this hypothesis is to correlate how well scene sounds were identified (see the Analysis section in the Supplemental Material), and the congruency effect ($r = -.15$, $p = .44$; Fig. 6). Although this is a rather indirect test of the influence of top-down knowledge, the lack of a correlation implies that explicit recognition of scene sounds was not necessary to produce the effect and is thus less likely to drive the general effect of sounds over vision. Consequently, we suggest that these audiovisual interactions between naturalistic sounds and visual objects may be driven by a lifetime of experience (Bey & McAdams, 2002; Körding et al., 2007), and the computational architecture that

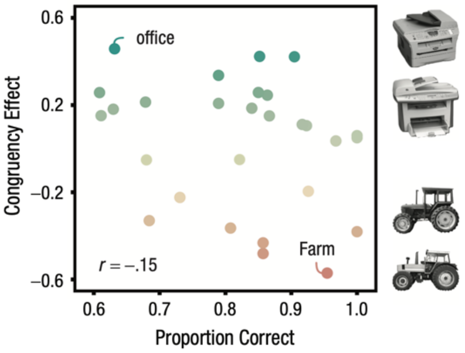


Fig. 6. Results from a preliminary survey. The survey revealed no reliable relationship between the identifiability of an auditory scene of the stimuli used in Experiment 1b (correct identification rate; *x*-axis) and the effect of congruent auditory scenes on response time in the visual-discrimination task (effect size of congruency from Experiment 1b; *y*-axis). Object images on the right side depict the stimuli used for the visual-discrimination phase of the two example sounds ("office" and "farm").

gives rise to them is an automatic feature of a probabilistic mind (de Lange et al., 2018; Kok et al., 2012; Versace et al., 2009). However, future studies should more directly test to what extent the present cross-modal effects occur effortlessly through relatively automatic integration of audiovisual signals, or to what extent volition and top-down control play a role.

Together, across four experiments and a replication, we demonstrated that processing of auditory objects and scenes facilitates visual perception of related visual objects, even when sounds are not predictive of the visual stimulus in a given task and even when participants have no prior experience with the particular set of audiovisual stimuli. Whether these effects systematically vary between different stimulus categories (animate vs. inanimate, natural vs. manmade, etc.) and generalize to novel categories should be explored in future studies. Collectively, our results suggest that perception integrates contextual information at various levels of processing and can leverage general, gist-like information across sensory modalities to facilitate visual object perception.

Transparency

Action Editor

Editor

Author Contributions

Jamal R. Williams: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Resources; Software; Validation; Visualization; Writing – original draft; Writing – review & editing.

Viola S. Störmer: Conceptualization; Funding acquisition; Investigation; Methodology; Project administration; Resources; Supervision; Visualization; Writing – review & editing.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This work was supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-2038238.

Open Practices

No aspects of Experiment 1a were preregistered; all materials, data, and analysis scripts are publicly available online (https://osf.io/msqnv/). For Experiment 1b, the design, sample size, exclusion criteria, and analysis plan were preregistered on AsPredicted (https://aspredicted.org/VP2_1VN), and all materials, data, and analysis scripts are publicly available online (https://osf.io/msqnv/).

No aspects of Experiment 2a were preregistered. All materials, data, and analysis scripts are publicly available online (https://osf.io/msqnv/). The design, sample size, exclusion criteria, and analysis plan for Experiment 2b were preregistered on AsPredicted (https://aspredicted.org/VFV_44W), and all materials, data, and analysis scripts are publicly available online (https://osf.io/msqnv/).

This article has received the badges for Open Data, Open Materials, and Preregistration. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

ORCID iD

Jamal R. Williams https://orcid.org/0000-0002-3034-511X

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/09567976241237737

References

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology*, *14*(3), 257–262. https://doi.org/10.1016/j.cub.2004.01.029

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience*, *5*(8), 617–629. https://doi.org/10.1038/nrn1476

Brandman, T., Avancini, C., Leticevscaia, O., & Peelen, M. V. (2020). Auditory and semantic cues facilitate decoding of visual object category in MEG. *Cerebral Cortex*, *30*(2), 597–606. https://doi.org/10.1093/cercor/bhz110

Chen, Y.-C., & Spence, C. (2010). When hearing the bark helps to identify the dog: Semantically-congruent sounds modulate the identification of masked pictures. *Cognition*, *114*(3), 389–404. https://doi.org/10.1016/j.cognition.2009.10.012

Chen, Y.-C., & Spence, C. (2011). Crossmodal semantic priming by naturalistic sounds and spoken words enhances visual sensitivity. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(5), 1554–1568. https://doi.org/10.1037/a0024329

Davenport, J. L., & Potter, M. C. (2004). Scene consistency in object and background perception. *Psychological Science*, *15*(8), 559–564. https://doi.org/10.1111/j.0956-7976.2004.00719.x

de Haas, B., Schwarzkopf, D. S., Urner, M., & Rees, G. (2013). Auditory modulation of visual stimulus encoding in human retinotopic cortex. *NeuroImage*, *70*, 258–267. https://doi.org/10.1016/j.neuroimage.2012.12.061

Draschkow, D., & Võ, M. L.-H. (2017). Scene grammar shapes the way we interact with objects, strengthens memories, and speeds search. *Scientific Reports*, *7*(1), Article 16471. https://doi.org/10.1038/s41598-017-16739-x

Edmiston, P., & Lupyan, G. (2015). What makes words special? Words as unmotivated cues. *Cognition*, *143*, 93–100. https://doi.org/10.1016/j.cognition.2015.06.008

Hellmann, S., Zehetleitner, M., & Rausch, M. (2023). Simultaneous modeling of choice, confidence, and response time in visual perception. *Psychological Review*, *130*(6), 1521–1543. https://doi.org/10.1037/rev0000411

Heron, J., Whitaker, D., & McGraw, P. V. (2004). Sensory uncertainty governs the extent of audio-visual interaction. *Vision Research*, *44*(25), 2875–2884. https://doi.org/10.1016/j.visres.2004.07.001

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal inference in multisensory perception. *PLOS ONE*, *2*(9), Article e943. https://doi.org/10.1371/journal.pone.0000943

Liu, B., Wu, G., & Meng, X. (2012). Cross-modal priming effect based on short-term experience of ecologically unrelated audio-visual information: An event-related potential study. *Neuroscience*, *223*, 21–27. https://doi.org/10.1016/j.neuroscience.2012.06.009

Lupyan, G., & Thompson-Schill, S. L. (2012). The evocative power of words: Activation of concepts by verbal and nonverbal means. *Journal of Experimental Psychology: General*, *141*(1), 170–186. https://doi.org/10.1037/a0024904

Lupyan, G., & Ward, E. J. (2013). Language can boost otherwise unseen objects into visual awareness. *Proceedings of the National Academy of Sciences, USA*, *110*(35), 14196–14201. https://doi.org/10.1073/pnas.1303312110

Meyerhoff, H. S., & Huff, M. (2016). Semantic congruency but not temporal synchrony enhances long-term memory performance for audio-visual scenes. *Memory & Cognition*, *44*(3), 390–402. https://doi.org/10.3758/s13421-015-0575-6

Mulatti, C., Treccani, B., & Job, R. (2014). The role of the sound of objects in object identification: Evidence from picture naming. *Frontiers in Psychology*, *5*, Article 1139. https://www.frontiersin.org/articles/10.3389/fpsyg.2014.01139

Norman, D. A., & Wickelgren, W. A. (1969). Strength theory of decision rules and latency in retrieval from short-term memory. *Journal of Mathematical Psychology*, *6*(2), 192–208. https://doi.org/10.1016/0022-2496(69)90002-9

Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences*, *11*(12), 520–527. https://doi.org/10.1016/j.tics.2007.09.009

Palmer, Stephen, E. (1975). The effects of contextual scenes on the identification of objects. *Memory & Cognition*, *3*(5), 519–526. https://doi.org/10.3758/BF03197524

Plass, J., Guzman-Martinez, E., Ortega, L., Suzuki, S., & Grabowecky, M. (2017). Automatic auditory disambiguation of visual awareness. *Attention, Perception, & Psychophysics*, *79*(7), 2055–2063. https://doi.org/10.3758/s13414-017-1355-0

Ratcliff, R., & Starns, J. J. (2013). Modeling confidence judgments, response times, and multiple choices in decision making: Recognition memory and motion discrimination. *Psychological Review*, *120*(3), 697–719. https://doi.org/10.1037/a0033152

Schneider, T. R., Engel, A. K., & Debener, S. (2008). Multisensory identification of natural objects in a two-way crossmodal priming paradigm. *Experimental Psychology*, *55*(2), 121–132. https://doi.org/10.1027/1618-3169.55.2.121

Spence, C. J., & Driver, J. (1994). Covert spatial orienting in audition: Exogenous and endogenous mechanisms. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(3), 555–574. https://doi.org/10.1037/0096-1523.20.3.555

Stein, T., & Peelen, M. V. (2015). Content-specific expectations enhance stimulus detectability by increasing perceptual sensitivity. *Journal of Experimental Psychology: General*, *144*(6), 1089–1104. https://doi.org/10.1037/xge0000109

Vetter, P., Bola, Ł., Reich, L., Bennett, M., Muckli, L., & Amedi, A. (2020). Decoding natural sounds in early "visual" cortex of congenitally blind individuals. *Current Biology*, *30*(15), 3039–3044.e2. https://doi.org/10.1016/j.cub.2020.05.071

Vetter, P., Smith, F. W., & Muckli, L. (2014). Decoding sound and imagery content in early visual cortex. *Current Biology*, *24*(11), 1256–1262. https://doi.org/10.1016/j.cub.2014.04.020

Welch, B. L. (1947). The generalization of 'student's' problem when several different population variances are involved. *Biometrika*, *34*(1/2), 28–38. https://doi.org/10.2307/2332510

Williams, J. R., Markov, Y. A., Tiurina, N. A., & Störmer, V. S. (2022). What you see is what you hear: Sounds alter the contents of visual perception. *Psychological Science*, *33*(12), 2109–2122. https://doi.org/10.1177/09567976221121348

Wilson, B. M., Harris, C. R., & Wixted, J. T. (2020). Science is not a signal detection problem. *Proceedings of the National Academy of Sciences, USA*, *117*(11), 5559–5567. https://doi.org/10.1073/pnas.1914237117

Winkler, I., Denham, S. L., & Nelken, I. (2009). Modeling the auditory scene: Predictive regularity representations and perceptual objects. *Trends in Cognitive Sciences*, *13*(12), 532–540. https://doi.org/10.1016/j.tics.2009.09.003

Acknowledgements

Chapter 2, in full, is a reprint of the material as it appears in Psychological Science, 2024, Williams, Jamal R.; Störmer, Viola S. The dissertation author was the primary investigator and author of this paper.

Chapter 3 What You See Is What You Hear: Sounds Alter the Contents of Visual Perception

Jamal R. Williams, Yuri A. Markov, Natalia A. Tiurina, and Viola S. Störmer

As it appears in

Psychological Science

2022

Volume 33(12), Pages 2109-2122

Research Article

ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Psychological Science
1–14
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/09567976221121348
www.psychologicalscience.org/PS
SAGE

What You See Is What You Hear: Sounds Alter the Contents of Visual Perception

Jamal R. Williams[1], **Yuri A. Markov**[2], **Natalia A. Tiurina**[2], **and Viola S. Störmer**[1,3]

[1]Department of Psychology, University of California San Diego; [2]Laboratory of Psychophysics, Brain Mind Institute, Ecole Polytechnique Federale de Lausanne (EPFL); and [3]Department of Brain and Psychological Sciences, Dartmouth College

Abstract

Visual object recognition is not performed in isolation but depends on prior knowledge and context. Here, we found that auditory context plays a critical role in visual object perception. Using a psychophysical task in which naturalistic sounds were paired with noisy visual inputs, we demonstrated across two experiments (young adults; *ns* = 18–40 in Experiments 1 and 2, respectively) that the representations of ambiguous visual objects were shifted toward the visual features of an object that were related to the incidental sound. In a series of control experiments, we found that these effects were not driven by decision or response biases (*ns* = 40–85) nor were they due to top-down expectations (*n* = 40). Instead, these effects were driven by the continuous integration of audiovisual inputs during perception itself. Together, our results demonstrate that the perceptual experience of visual objects is directly shaped by naturalistic auditory context, which provides independent and diagnostic information about the visual world.

Keywords

visual perception, crossmodal perception, multisensory integration, object recognition, open data, open materials

Received 1/25/22; Revision accepted 8/1/22

When we look around the world, pertinent visual information can be ambiguous or indeterminate. To overcome this problem and to form meaningful representations, the visual system not only relies on the visual features of an object itself but also incorporates prior knowledge and concurrently available contextual information (Bar, 2004; Biederman et al., 1973; Davenport & Potter, 2004). This integration of available information is not exclusively visual either, as available information from every sensory system is evaluated, weighed, and integrated to form a complete perceptual experience (Alais & Burr, 2004; Chen & Spence, 2010, 2011a; Ernst & Banks, 2002; Körding et al., 2007; Schneider et al., 2008). However, most of the work on multisensory integration has focused on characterizing how hearing a sound can *facilitate* visual processing; here, we investigated whether naturalistic sounds alter our phenomenology of visual objects. In other words, does the sound of a seal barking change our visual experience and make visual information appear more seal-like than it actually is? Or do sounds simply improve perceptual processing of related visual objects by speeding responses or improving accuracy.

It is well established that simple auditory information, such as a noise burst or a beep, can influence visual processing of low-level visual stimuli quite dramatically, for example by enhancing their early visual processing (Giard & Peronnet, 1999; McDonald et al., 2000; Störmer et al., 2009; Vroomen & De Gelder, 2000) or by disambiguating visual motion stimuli (Sekuler et al., 1997; Watanabe & Shimojo, 2001). Naturalistic sounds have also been found to affect higher-level visual processing, such that response times (RTs) are faster and accuracy is higher in object recognition tasks when sight and sound are congruent relative to incongruent (Chen &

Corresponding Author:
Jamal R. Williams, Department of Psychology, University of California San Diego
Email: jrwilliams@ucsd.edu

Spence, 2011a; Williams & Störmer, 2019). However, it is unclear whether real-world sounds simply enhance perceptual processing—leading to a more rapidly achieved or more accurate representation for congruent audiovisual conditions—or whether sounds can change how we see visual objects. Here, we focused on testing this hypothesis by investigating whether incidental naturalistic sounds can alter the visual representations of pertinent visual objects.

We addressed these questions by investigating how naturalistic sounds modulate the visual processing of ambiguous objects. We used a visual discrimination task with a perceptual locus (Sadr & Sinha, 2004; Williams & Störmer, 2019) and designed a novel set of object stimuli that were paired at random with related or unrelated sounds. Because the influence of sound on vision seems particularly effective when visual information is noisy or dubious—where sounds provide independent and unequivocal clues about the visual environment (Alais & Burr, 2004; Heron et al., 2004; Rohe & Noppeney, 2015; Watanabe & Shimojo, 2001)—we used ambiguous visual stimuli paired with clear and distinct sounds. Specifically, we created a set of ambiguous visual stimuli by morphing together the features of two visual objects (objects A and B, e.g., a hammer and a seal; Fig. 1a) and presented these stimuli with naturalistic sounds that were congruent with one of these progenitor objects. Visual objects and sounds were presented simultaneously, and participants looked for a target object in visual noise, after which they precisely reported that object using a continuous report method. We examined whether participant's reports of the visual objects were altered by the sounds they heard—in particular, whether sounds would shift the perceptual representation toward the features related to the sound. In a series of control experiments, we also tested at what processing stage these audiovisual effects arose and found evidence consistent with the hypothesis that the effects of sounds on visual object recognition have an early, perceptual locus.

All data, scripts, and stimuli needed to replicate these experiments and analyses are available on OSF (https://osf.io/85kwv).

Experiment 1

On each trial, an ambiguous visual stimulus that was a morph of two objects (i.e., the target morph; see Fig. 1) slowly faded into view from visual noise, while the sound of a real-world object played. Participants were instructed to press a button as soon as they could accurately recreate the target morph using continuous report (Fig. 1b), in which they had to adjust a test object to the one they had seen during the visual discrimination phase as accurately as possible. Critically, the sounds could be either related or unrelated to the target morph: Unrelated sounds were highly dissimilar from the target morph (e.g., a whistling train for the hammer–seal morphs), whereas related sounds matched the identity of one of the target morph's anchor objects.[1]

Statement of Relevance

Perception is inherently multisensory, and even senses that might appear to be irrelevant play a role in how we perceive the world. To what extent do our senses influence and change our perceptual experience? For example, imagine you catch a glimpse of something rapidly flying by a window. Because it could be any number of things, auditory information could be incredibly useful for resolving this uncertainty: A buzzing would suggest it was a drone, whereas a caw suggests it was a crow. Does the sound of a drone make this dubious object appear more drone-like than it otherwise would have? Here, we tested how naturalistic sounds affect the perception of visual objects and found that object representations are shifted toward the visual features that are congruent with the sound. These findings demonstrate that what we hear has profound impacts on how we perceive the visual world.

Method

Participants. All participants gave informed consent in accordance with the procedures approved by the institutional review board at the University of California (UC) San Diego. Participants were between 18 and 25 years old and reported having normal hearing and normal or corrected-to-normal vision. Twenty-five undergraduates (14 women; mean age = 20.6 years) from UC San Diego took part in our online Experiment 1a in exchange for course credit. Data from six participants were removed because of poor task performance, leaving 19 participants in the final sample (see the Analysis section for more details on exclusion criteria). In Experiment 1b, 49 undergraduates (35 women; mean age = 20.52 years) from UC San Diego took part in this online study in exchange for course credit. Data from nine participants were removed because of poor task performance, leaving 40 participants in the final sample. Experiment 1b included more participants to ensure that we could detect potentially smaller effects after shortening the experiment to make it more suitable for online testing. To determine an appropriate number of participants, we

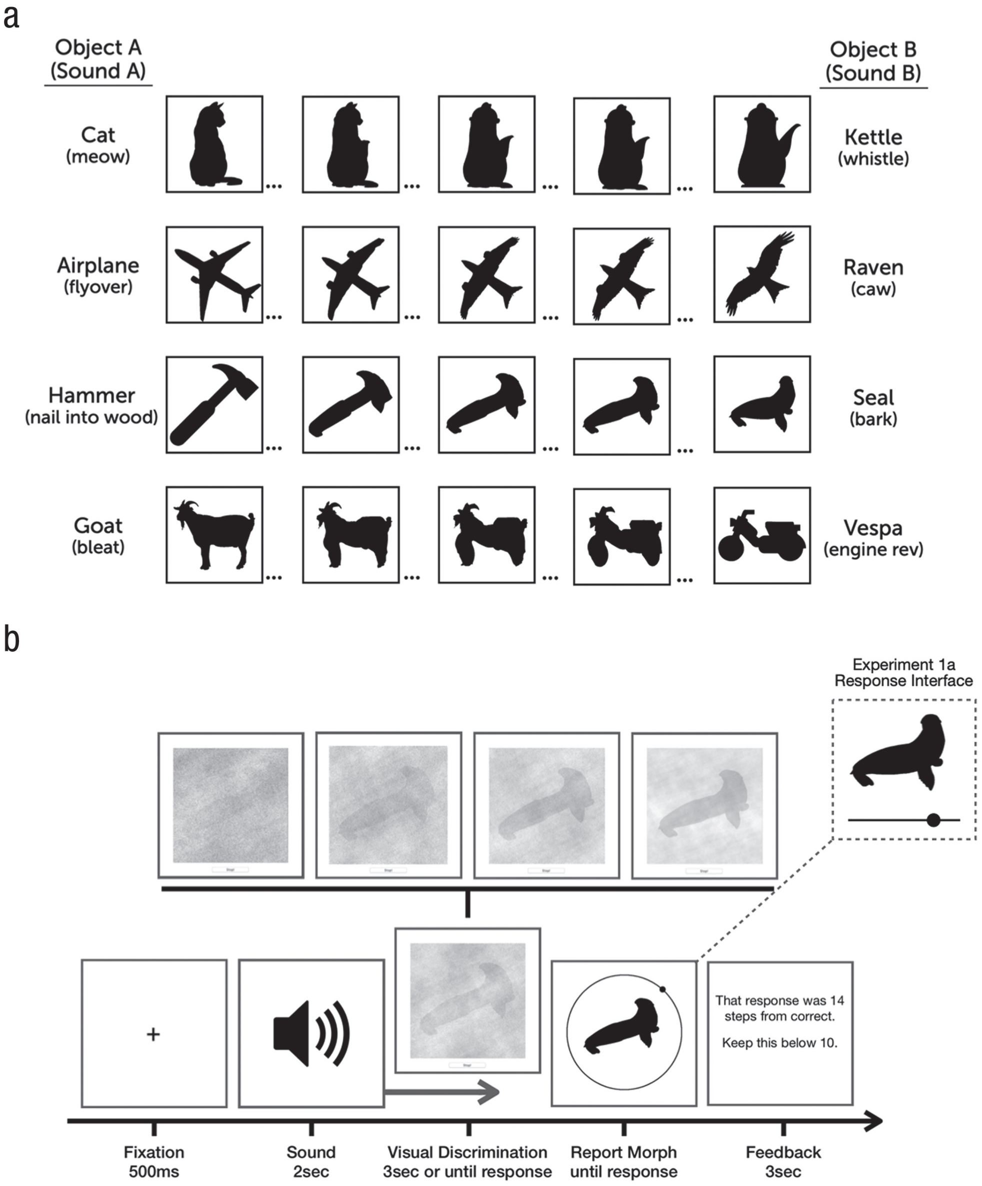


Fig. 1. Stimuli and task. (a) The four object pairs used in the experiments. The leftmost column shows anchor objects A, and the rightmost column shows anchor objects B (anchor-object sounds are shown in parentheses). Between each anchor object were 98 unique morphs that maintained features of both anchor objects. (b) General task design. Sounds played while a noisy object slowly faded into view (an example of the denoising process is shown above the visual discrimination panel). Experiment 1a used a linear response slider, whereas Experiment 1b used a circular response wheel.

performed a power analysis on the data obtained from Experiment 1a and found that we could adequately detect effects of sound on report error as low as Cohen's d_z = 0.45 with a sample size of 40 (power of .8 and an α of .05; using the base R package *pwr*; Champley et al., 2018).

Stimuli. A total of 12 real-world sounds were selected from online repositories (e.g., BBC Sound Effects, bbcsfx.acropolis.org.uk; freesound, freesound.org) and were edited to be 2 s in length and have roughly equivalent amplitudes (within and across stimuli when played at roughly 70 dB sound pressure level). Eight of the sounds were paired such that each sound in that pair could be distinct on the basis of auditory and semantic qualities (e.g., hammer–seal; see Fig. 1a). For each object pair, we collected and edited an additional unique sound that was unrelated to the audiovisual object pair. Unrelated sounds were selected to be as distinct as possible from the object pair (e.g., a train whistle for the hammer–seal object pair), whereas related sounds were selected to closely match sounds made by either anchor object A or anchor object B (see Fig. 1a for object sounds). For each sound pair, we collected or created a silhouette of a visual object that matched the object identity of the sound. For ambiguous objects, each silhouette also needed to share visual similarities such as shape, contours, and orientation with the silhouette from the other side of the object continuum. Using each silhouette as end points, we generated a set of 100 novel silhouettes by morphing the features of the two objects (object A and object B) for each object pair (Fig. 1a). We used a morphing program to fuse objects together and create these ambiguous morph pairs (Liao et al., 2014). The morphing procedure optimizes the retention of the original image features while avoiding ghosting artifacts and is based on three principal parameters: similarity (to match regions of images with similar edge structure), smoothness of the mapping (resulting vector fields favor the affine function in the absence of other constraints), and deviation from user-specified correspondence. We manually added specified correspondence points to resolve ambiguities and increase morphing performance.

Because the morphing process creates relatively arbitrary, psychologically nonuniform steps between 1 and 100, individual morph steps were rated in a separate online study to assess the psychometric functions for each of the morph pairs and to measure how the physical morph steps related to perceptual similarity. Here, participants were shown object A and object B (the unique images that anchored the end points of the continuum) and reported whether a test morph (randomly selected from the continuum) was visually more similar to object A or B. From these data, we generated psychophysical curves and selected three morphs from each object-pair continuum that corresponded to the points where 20%, 50%, and 80% of responses indicated that the morph appeared more as object A relative to object B. Note that although we aimed to introduce variability and greater ambiguity in the target stimulus set by selecting three different steps for each object pair, we planned to collapse data across these different morph levels for our main analysis to obtain adequate power. In sum, the image set contained four unique object pairs, each with three unique morphs (12 images total).

For the visual discrimination phase, stimuli were edited to form a continuous and difficult perceptual task that would allow the simultaneous presentation of a sound and a noisy visual object. First, to create noise masks that would effectively obscure the target silhouettes, we combined all 12 silhouette images and completely randomized the phase of this composite image. Thus, the power spectrum of the resulting noise image was correlated with that of all silhouettes and was completely unrecognizable. Then, we created a simple random noise mask using the function imnoise() in MATLAB (The MathWorks, Natick, MA) and overlaid this random noise mask on top of the phase-scrambled noise mask. Together, this resulted in a mask that effectively obscured the target morph silhouettes with both phased and random noise (see Fig. 1b).

Throughout each trial, the mask slowly became more transparent to reveal more of the underlying target image until only 40% of the noise mask remained. Also, on each trial, the phase of the target image was initially randomized 100% and then faded into a recognizable morph by slowly reducing the phase randomization until it was fully intact. The exact parameters of how quickly the noise faded and the target morph became more visible were based on pilot data from an in-lab version, which showed that participants could recognize the image when 60% image clarity was reached, which took roughly 3 s. All phase randomizations and noise masks were created prior to the online experiment; this ensured that the exact same stimuli were viewed by each participant.

Procedure. Participants performed 240 or 120 trials (Experiment 1a and 1b, respectively) that were split among three sound conditions: 40% of the sounds were related to visual anchor object A (e.g., the sound of hammering a nail into wood), and another 40% were related to visual anchor object B (e.g., the sound of a seal barking), and the remaining 20% served as a baseline condition, were unrelated to the visual object pair, and did not match either of the anchor objects. The nonmatching, unrelated sounds were selected to be unrelated to either of the sounds or visual objects. Related sounds were not predictive of which target morph appeared as the target in the visual discrimination phase (e.g., the sound of a seal barking could be presented when any of the three target morphs were presented). Each trial began with the playback of a 2-s sound of a real-world object, and participants were instructed to attend to this sound. Five hundred milliseconds after the sound onset, the visual discrimination phase appeared centrally (400 × 400 pixels) on the participant's browser of choice. The visual object always started completely obscured by visual noise and would slowly fade in to become more visible as time elapsed. More specifically, visual noise levels decreased by roughly 1% every 50 ms until the participant clicked the mouse to indicate that they had enough visual information to accurately perform the subsequent continuous report of the target object.

The mouse click stopped the visual discrimination phase, and if participants did not press the button within 3 s—when the phase randomization reached 40% noise and the object was identifiable though still obscured by noise—they received feedback encouraging them to accumulate visual information more quickly (these trials were discarded and not analyzed). Target images were randomly chosen on each trial and paired with one of the three sound conditions. Once the visual discrimination phase was completed, participants were presented with the response interface: A response silhouette (300 × 300 pixels) was shown as a probe above a continuous response slider (400 pixels wide). The probe was chosen randomly from the possible morph steps (1–100), and participants clicked and dragged a response dot along the continuous response line until they matched the probe to the target morph from the visual discrimination phase. Participants locked their response by clicking the mouse and then received feedback on their error (number of steps from the correct answer for 3 s). Participants then clicked to initiate the next trial.

Experiment 1a used a linear response slider in which the leftmost edge corresponded to anchor object A (Morph Step 1) and the rightmost edge corresponded to anchor object B (Morph Step 100). Further, we used three distinct morphs per object pair, and these morphs corresponded to three similar positions on the response slider across trials. Thus, it is possible that participants used these reliable positions along the response slider as a cue when responding—instead of focusing on the visual features of the response morph itself. To mitigate these concerns, and to replicate the effects of Experiment 1a using a different response format, in Experiment 1b, we presented participants with a response wheel that was rotated randomly on every trial so that there was no correspondence between positions on the response wheel and the visual response morph presented centrally, across trials (see Fig. 1b). Thus, the task in Experiment 1b was identical to that in Experiment 1a except that participants performed only half of the trials (and thus had less exposure and practice with these stimuli and task) and when the response screen appeared, a black ring (400 × 400 × 3 pixels) with a small position dot (50 × 50 pixels) surrounded the response morph (300 × 300 pixels). On every trial, the response ring was rotated by a random amount so that the angle of the position dot corresponded with a distinct morph step across trials. Thus, participants were not able to use the response interface itself as an anchor to find a particular morph but had to solely rely on the response morph, which was changing continuously as participants moved along the response wheel.

Analysis. For each sound condition (unrelated or related: A and B), we calculated a participant's median RT on the visual discrimination phase and their mean report error on the continuous report phase by sound condition. When comparing RT, we first checked to see whether RT differed between related sound A and sound B conditions. Across all experiments, we found no difference and thus collapsed RT estimates across sound A and B when comparing related and unrelated conditions. Error on continuous report was determined as the number of morph steps between the correct response (target morph) and the provided response. Morphs were numbered 1 to 100, and negative responses represent a response that is closer to 1 (object A) than the correct response and vice versa for positive responses. We calculated a participant's mean error per sound condition (sound A, B, and unrelated) and submitted these data to an analysis of variance (ANOVA). Report error in each figure is represented as the difference in average error between the related and unrelated conditions.

Exclusion criteria were decided in advance on the basis of preliminary pilot data. Data from participants were excluded if their average report error or average RT exceeded 3 standard deviations from the group mean. Furthermore, for each individual participant, all trials on which report error or RT exceeded 4 standard deviations from their mean were excluded. Last, any trials on which participants did not respond in the visual discrimination phase—instead opting to wait the entire duration of the trial—were excluded from further analysis. Data from participants were excluded from further analysis if more than 10% of trials were missing from their data set because of this removal process.

Results

We first submitted report error and RT on the visual discrimination and the continuous report phases, respectively, to a multivariate ANOVA (MANOVA). Here, we found a significant effect of sound on both RT and error, $F(4, 72) = 6.67$, $p < .001$, $\eta^2 = .27$. We next submitted continuous report responses to an ANOVA, which revealed a significant main effect of sound (sound A, sound B, or an unrelated sound), $F(2, 36) = 10.05$, $p < .001$, $\eta^2 = .36$. Our primary interest was whether related sounds A and B affected the same visual stimulus differently; thus, we next compared the mean error for each related sound with the error on unrelated sound trials—which matched the complexity and naturalistic properties of the related sounds, thus effectively serving as a neutral condition. These subsequent pairwise comparisons revealed that the sounds corresponding to anchor object A shifted responses toward that side of the object-morph continuum and away from responses on unrelated trials, $t(18) = -2.16$, $p = .044$, Cohen's $d_z = 0.50$, 95% confidence interval (CI) = [0.031, 0.959], whereas sounds corresponding to object B shifted responses in the opposite direction with roughly equal magnitude, $t(18) = 2.57$, $p = .019$, Cohen's $d_z = 0.59$, 95% CI = [0.199, 0.979] (see Fig. 2b). We next focused on RT during visual discrimination, which reflected the rate at which visual information was meaningfully integrated into a complete object. Participants were faster, on average, when they heard a related sound (1,638 ms) compared with an unrelated sound (1,682 ms), $t(18) = 2.47$, $p = .023$, Cohen's $d_z = 0.57$, 95% CI = [0.198, 0.936]. This difference suggests that, on unrelated trials, participants required roughly 10% more visual evidence than on related trials to perform the task with roughly equal levels of accuracy (mean absolute error = 6.00 vs. 6.07), $t(18) = 0.39$, $p = .7$, Cohen's $d_z = 0.089$, 95% CI = [0.086, 0.093], Bayes factor favoring the null over the alternative hypothesis (BF_{01}) = 3.94. Thus, auditory information accelerated visual feature extraction from the noisy images and possibly increased participants' confidence in their visual judgments as well (Williams & Störmer, 2019). Additionally, we conducted a linear mixed-effects analysis to account for variability in the stimulus set and, after accounting for this variance, found a main effect of sound for RT, $\chi^2(2) = 6.27$, $p = .043$, and report error, $\chi^2(2) = 6.05$, $p = .048$.

Experiment 1a used a linear response interface where the leftmost edge corresponded to anchor object A (Morph Step 1) and the rightmost edge corresponded to anchor object B (Morph Step 100). It is therefore possible that participants used these reliable positions along the response slider as a cue when responding—instead of focusing on the visual features of the response morph itself. To mitigate these concerns in Experiment 1b, we implemented a response wheel that rotated randomly on every trial (Fig. 1b). We submitted RT and report error to a MANOVA and found a main effect of sound, $F(4, 156) = 8.508$, $p < .001$, $\eta^2 = .18$. Next, we found that sounds had a reliable effect on report error, $F(2, 78) = 11.23$, $p < .001$, $\eta^2 = .22$, and that related sounds shifted responses away from the average error on unrelated trials and toward the visual features of anchor object A, $t(39) = -2.58$, $p = .014$, Cohen's $d_z = 0.41$, 95% CI = [–0.20, 1.02], and object B, $t(29) = 2.77$, $p = .01$, Cohen's $d_z = 0.42$, 95% CI = [–0.31, 1.16] (Fig. 2). RT from the visual discrimination phase was again faster when sounds were related to the target morph (M = 1,798 ms) than when they were unrelated (M = 1,860 ms), $t(39) = 3.22$, $p = 0.003$, Cohen's $d_z = 0.51$, 95% CI = [0.12, 0.89], and like before, this difference in RT did not result in a reliable difference in accuracy (Ms = 6.83 vs. 6.84), $t(39) = 0.03$, $p = .98$, Cohen's $d_z = 0.004$, 95% CI = [–0.001, 0.009], $BF_{01} = 5.86$. We also found that the variability in our stimulus set did not extinguish the main effect of sound on RT, $\chi^2(2) = 29.44$, $p < .001$, or report error, $\chi^2(2) = 29.44$, $p < .001$. Taken together, the results from these experiments demonstrate that related auditory information speeds visual object processing while also shifting feature representations of visual objects toward those features that match the incidental auditory context.

Experiment 2

The results of Experiments 1a and 1b led us to hypothesize that sounds influence concurrent visual processing by shifting ambiguous visual inputs toward visual features that are congruent with the sound. However, it could be that sounds influence later, nonperceptual processing stages, such as decisional and response processes. Although such a postperceptual account seems incompatible with faster RT for related relative to unrelated sounds, we directly tested this alternative in Experiments 2a and 2b by presenting sounds when they should have the greatest impact over decisional processes: during the continuous report phase.

Method

Participants. In Experiment 2a, all participants were 18 to 23 years old (mean age = 20.36 years), reported normal hearing and normal or corrected-to-normal vision, and gave informed consent in accordance with the procedures approved by the institutional review board at UC San Diego. Forty-nine participants (32 women) from UC San Diego took part in this online experiment in exchange for course credit. Data from seven participants were removed using the same criteria as described above, leaving 40 participants in the final sample. In Experiment 2b, all participants were 18 to 34 years old (mean age = 20.69 years), and 96 participants (76 women) from UC

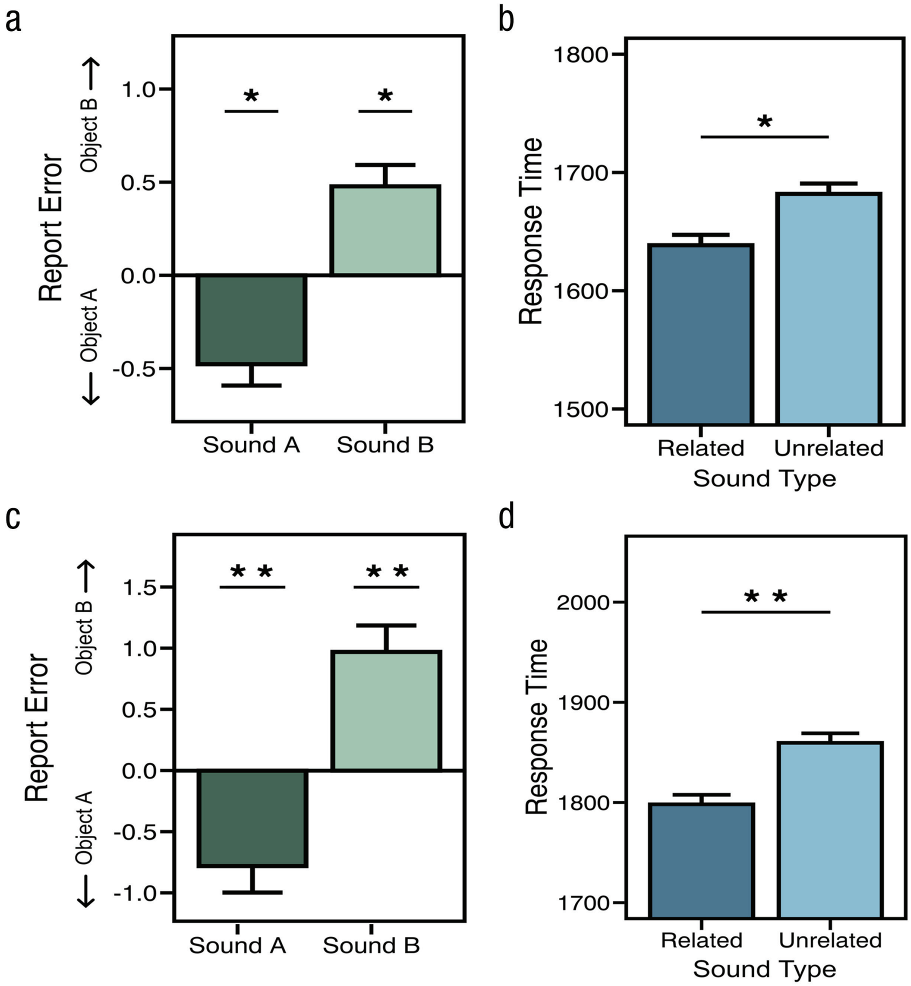


Fig. 2. Data from Experiments 1a (top row) and 1b (bottom row). Average report error (difference from unrelated sounds) for Experiment 1a (a) and Experiment 1b (c) shows that related sounds influenced report error such that the response morph appeared more like the sound's anchor-object identity. The right column demonstrates that, for both Experiment 1a (b) and Experiment 1b (d), sounds influenced response time such that participants were faster when they heard a related sound compared with an unrelated one.
Error bars are +/-1 SEM, $*p < .05$. $**p < .01$. $***p < .001$.

San Diego took part in this online experiment in exchange for course credit. Data from 11 participants were removed using the same criteria as described above, leaving 85 participants in the final sample. Exclusion criteria were identical to those in Experiment 1.

Procedure. The task was identical to that in Experiment 1b, except that sounds now started to play immediately following the visual discrimination phase and during continuous report. Each trial began with the same visual discrimination phase, except with no sound and, after a button press, the visual input stopped, a real-world sound began to play, and the continuous report interface was presented (after 500 ms; Fig. 3). If the effect is largely driven by a decisional process (such as response bias or low-confidence responses), we would expect a similar, or perhaps even larger, effect of sound on visual perception relative to that found in Experiments 1a and 1b. If, however, real-world sounds primarily affect perceptual and not decisional processes, then this manipulation should eliminate or reduce the effect because perceptual processing is likely complete by the time participants begin reporting the target item. In Experiment 2b, on half of all trials, a sound started playing shortly before the visual discrimination task (as in Experiments 1a and 1b), and on the remaining half of the trials, the sound was played after the visual discrimination task and during the continuous response task (as in Experiment 2a). These sound-onset conditions were presented in blocks (30 trials per block) and interleaved.

Results

In Experiment 2a, we submitted RT and report error to a MANOVA and found no main effect of sound, *F*(4, 156) = 0.44, *p* = .78, η^2 = .01. Following up, we found that sounds had no significant impact on report error, *F*(2, 78) = 0.38, *p* = .69, η^2 = .009 (Fig. 3b), and as expected, RT on related (*M* = 1,911 ms) and unrelated trials (*M* = 1,906 ms) was not significantly different, *t*(39) = 0.29, *p* = .77, Cohen's d_z = 0.04, 95% CI = [−0.379, 0.286], BF_{01} = 5.63 (Fig. 3d). A closer analysis of report error found no significant impact of sound. Report error on neither sound A trials, *t*(39) = 0.82, *p* = .42, Cohen's d_z = 0.12, 95% CI = [−0.362, 0.621], nor sound B trials,

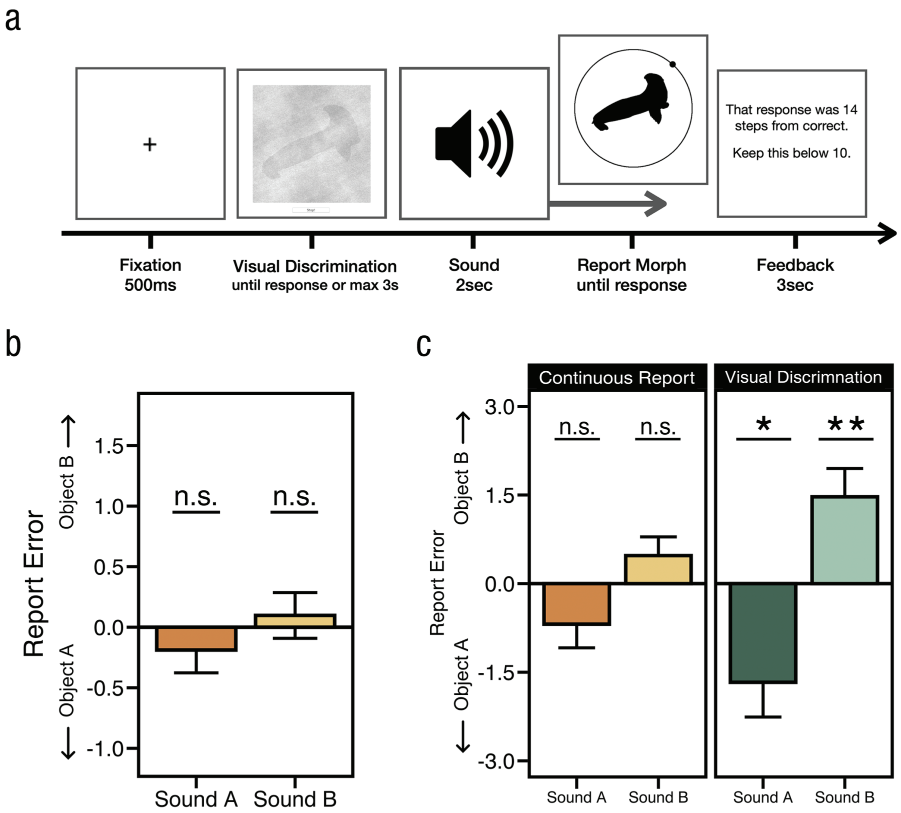


Fig. 3. *(continued on next page)*

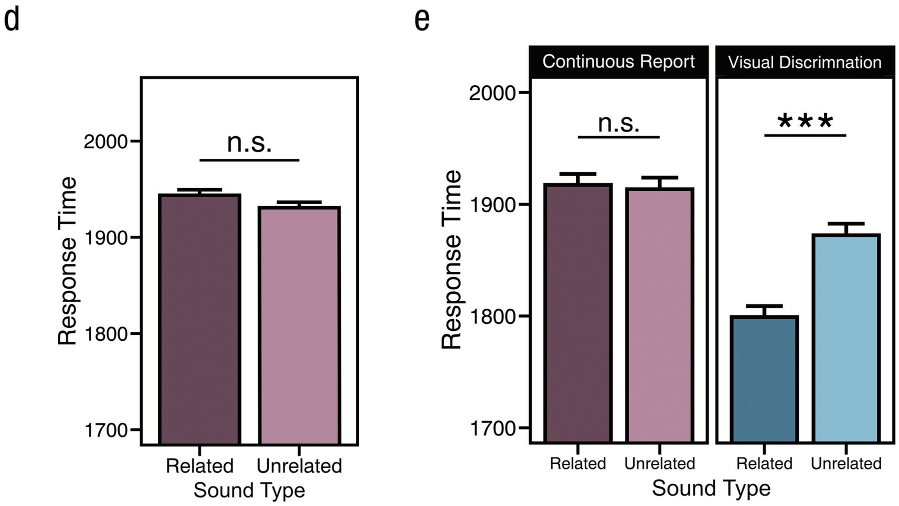


Fig. 3. Results and task design from Experiments 2a and 2b. (a) Task design. Sounds were always played during the continuous report phase in Experiment 2a and on half of all blocks in Experiment 2b. (b, c) Average report error (difference from unrelated) for Experiment 2a and Experiment 2b (separated by when the sound onset began). Related sounds influenced report error such that the response morph appeared more like the sound's anchor object when the sound was played during the visual discrimination phase (c, green bars) and not when played during the continuous report phase (b, c, orange bars). (d) Response time for related and unrelated trials in Experiment 2a. (e) Response time for related and unrelated trials in Experiment 2b, separated by when the sound was played: during continuous report (purple bars) or visual discrimination phases (blue bars). Results show that RT was reliably affected only when sounds were heard during the visual discrimination phase.
Error bars are +/-1 SEM, **p* < .05. ***p* < .01. ****p* < .001.

$t(39) = 0.24$, $p = .81$, Cohen's $d_z = 0.03$, 95% CI = [–0.831, 0.755], was significant, and we found compelling evidence to support these null findings ($BF_{01} = 4.28$ and 5.70, respectively).

In Experiment 2b, we combined manipulations from Experiments 1b and 2a in a within-subject design and varied whether sounds were played during the continuous report phase (as in Experiment 2a) or were played during the visual discrimination phase (as in Experiments 1a and 1b). We submitted RT and report error to a MANOVA with both sound and sound-onset conditions and found main effects of sound, $F(4, 336) = 9.19$, $p < .001$, $\eta_p^2 = .10$, and sound onset, $F(2, 83) = 6.16$, $p = .003$, $\eta_p^2 = .13$, as well as a significant interaction, $F(4, 336) = 4.76$, $p < .001$, $\eta_p^2 = .05$. We next focused on report error and found a main effect of sound, $F(2, 84) = 11.31$, $p < .001$, $\eta_p^2 = .12$; there was no main effect of sound onset (during or after visual discrimination), $F(1, 84) = 0.16$, $p = .69$, $\eta_p^2 = .001$, and there was a significant interaction, $F(2, 168) = 3.39$, $p = .036$, $\eta_p^2 = .04$. To explore the interaction, we compared the effect of sound on report error and found that sounds produced a significantly larger effect when they were played during the visual discrimination phase compared with when they were played during the continuous report phase, $t(84) = 2.34$, $p = .021$, Cohen's $d_z = 0.25$, 95% CI = [–1.425, 1.934] (see Fig. 3c).

We next analyzed report error independently for each sound-onset condition. When participants heard sounds during the visual discrimination phase, we found that related sounds shift responses toward anchor object A, $t(84) = 2.30$, $p = .024$, Cohen's $d_z = 0.25$, 95% CI = [–1.696, 1.198], and object B, $t(84) = 2.96$, $p = .004$, Cohen's $d_z = 0.32$, 95% CI = [–0.668, 1.309]. These results were significant after analyses accounted for stimulus variability as well, $\chi^2(2) = 58.59$, $p < .001$. However, and in contrast to these findings, when participants heard sounds during the continuous report phase (Fig. 3d), we found that error on unrelated trials was not significantly different from error on sound A trials, $t(84) = 1.56$, $p = 0.12$, Cohen's $d_z = 0.16$, 95% CI = [–1.043, 0.705], $BF_{01} = 2.61$, and sound B trials, $t(84) = 1.42$, $p = .16$, Cohen's $d_z = 0.15$, 95% CI = [–0.51, 0.818], $BF_{01} = 3.18$.

We then examined RT and found significant main effects of sound, $F(2, 168) = 7.34$, $p < .001$, $\eta_p^2 = .8$, and sound onset, $F(1, 84) = 12.25$, $p = .001$, $\eta_p^2 = .13$, as well as a significant interaction, $F(2, 168) = 6.12$, $p = .003$,

$\eta_p^2 = .07$. Participants were significantly faster on related trials (M = 1,779 ms) compared with unrelated trials (M = 1,852 ms), $t(84) = 4.05$, $p < .001$, Cohen's $d_z = 0.44$, 95% CI = [0.079, 0.7996] (Fig. 3e, blue bars), when sounds played during the visual discrimination phase, and this difference in RT did not lead to significant differences in accuracy (Ms = 7.76 vs. 7.31), $t(84) = 1.21$, $p = .23$, Cohen's $d_z = 0.13$, 95% CI = [−0.875, 0.612], $BF_{01} = 4.13$ (see Fig. 3e). As expected, we observed no significant difference in RT between the related (M = 1,899 ms) and unrelated (M = 1,903 ms) conditions when sounds were played during the continuous report phase, $t(84) = 0.25$, $p = .80$, Cohen's $d_z = 0.03$, 95% CI = [−0.287, 0.343], $BF_{01} = 7.99$ (Fig. 3e, purple bars).

Overall, RT was on average slower when sounds were played during the continuous report phase compared with the visual discrimination phase, but this difference in RT (i.e., having target images with lower levels of noise) was not statistically significant, $t(84) = 1.58$, $p = .12$, Cohen's $d_z = 0.17$, 95% CI = [−0.817, 0.475], $BF_{01} = 2.55$, and the numerical difference in RT did not lead to a significant difference in report error across sound-onset conditions (Ms = 7.38 vs. 7.54), $t(84) = 0.62$, $p = .54$, Cohen's $d_z = 0.06$, 95% CI = [−0.85, 0.725], $BF_{01} = 6.94$. These results replicate those of the previous experiments and demonstrate that sounds have their greatest influence when they are presented concurrently with visual information and can thus be integrated directly with incoming visual information.

Experiment 3

Experiments 2a and 2b suggest that this perceptual shifting is not largely driven by postperceptual mechanisms. However, another possibility is that the semantic content of these naturalistic sounds drives preperceptual, top-down influences on visual perception (although, top-down mechanisms may diminish multisensory effects). To test whether sounds might activate high-level semantic representations—that subsequently influence sensory processing—in Experiment 3, we presented the full length of a sound prior to the onset of the visual discrimination phase (cf. Cox & Hong, 2015; Lupyan & Ward, 2013), which provides the same audiosemantic content as before but should primarily drive preperceptual mechanisms that have been shown to require a longer delay between sound and target onset (Boutonnet & Lupyan, 2015; Chen & Spence, 2018a, 2018b; Lupyan & Ward, 2013).

Method

Participants. All participants were between 18 and 25 years old (mean age = 20.1 years), reported normal or corrected-to-normal vision, and gave informed consent in accordance with the procedures approved by the institutional review board at UC San Diego. Forty-eight undergraduates (25 women) from UC San Diego took part in this online study in exchange for course credit. Data from eight participants were removed using the same criteria as described above, leaving 40 participants in the final sample.

Procedure. The task was identical to that in Experiment 1b, except that sounds now preceded the visual discrimination task by 3 s (Fig. 4a). Each trial started with a real-world sound (2 s) and after it finished, the presentation of the visual discrimination task automatically began after the 3-s delay.

Results

We submitted RT and report error to a MANOVA and found no main effect of sound, $F(4, 156) = 2.18$, $p = .07$, $\eta^2 = .05$. We found that sounds did not have a significant impact on report error, $F(2, 78) = 2.08$, $p = .13$, $\eta^2 = .05$ (Fig. 4b), and we did not find a significant RT benefit for related sounds (Ms = 2,008 vs. 2,037 ms), $t(39) = 1.73$, $p = .09$, Cohen's $d_z = 0.27$, 95% CI = [−0.073, 0.619], $BF_{01} = 1.50$ (Fig. 4c). Preplanned t tests of report error further demonstrated that error on unrelated trials was not significantly different from error on sound A trials, $t(39) = 1.19$, $p = 0.24$, Cohen's $d_z = 0.18$, 95% CI = [−0.593, 0.968], $BF_{01} = 3.04$, or sound B trials, $t(39) = 0.64$, $p = .53$, Cohen's $d_z = 0.10$, 95% CI = [−0.725, 0.927], $BF_{01} = 4.84$. We also compared the small effect of sound that we found in Experiment 3 with that found in Experiment 1b and submitted error to an ANOVA. Here, we found no main effect of experiment, $F(1, 78) = 0.18$, $p = 0.67$, $\eta_p^2 < .01$), a main effect of sound, $F(1, 78) = 24.04$, $p < .0001$, $\eta_p^2 = .24$, and a significant interaction, $F(1, 78) = 4.21$, $p = .0435$, $\eta_p^2 = .05$.

Thus, the effects observed in Experiment 1b were above and beyond the small (and unreliable) effect observed in Experiment 3. Furthermore, the observed effect sizes across experiments, further support this: The average effect size (d_z) of report error for Experiment 3 was 0.13, and for Experiments 1a, 1b, and 2b, effect sizes (d_zs) ranged from 0.3 to 0.59. Overall, these results suggest that the effects we observed in Experiments 1a and 1b and Experiment 2b were largely driven by the continuous presentation of sight and sound and less so by attentional mechanisms or other top-down goals and expectations. This reinforces previous findings stressing the importance of the temporal overlap of incoming audiovisual stimuli, as predicted by multisensory integration accounts (Chen & Spence, 2011b, 2018a; Colonius & Diederich, 2004; Meredith et al., 1987; van Atteveldt et al., 2007).

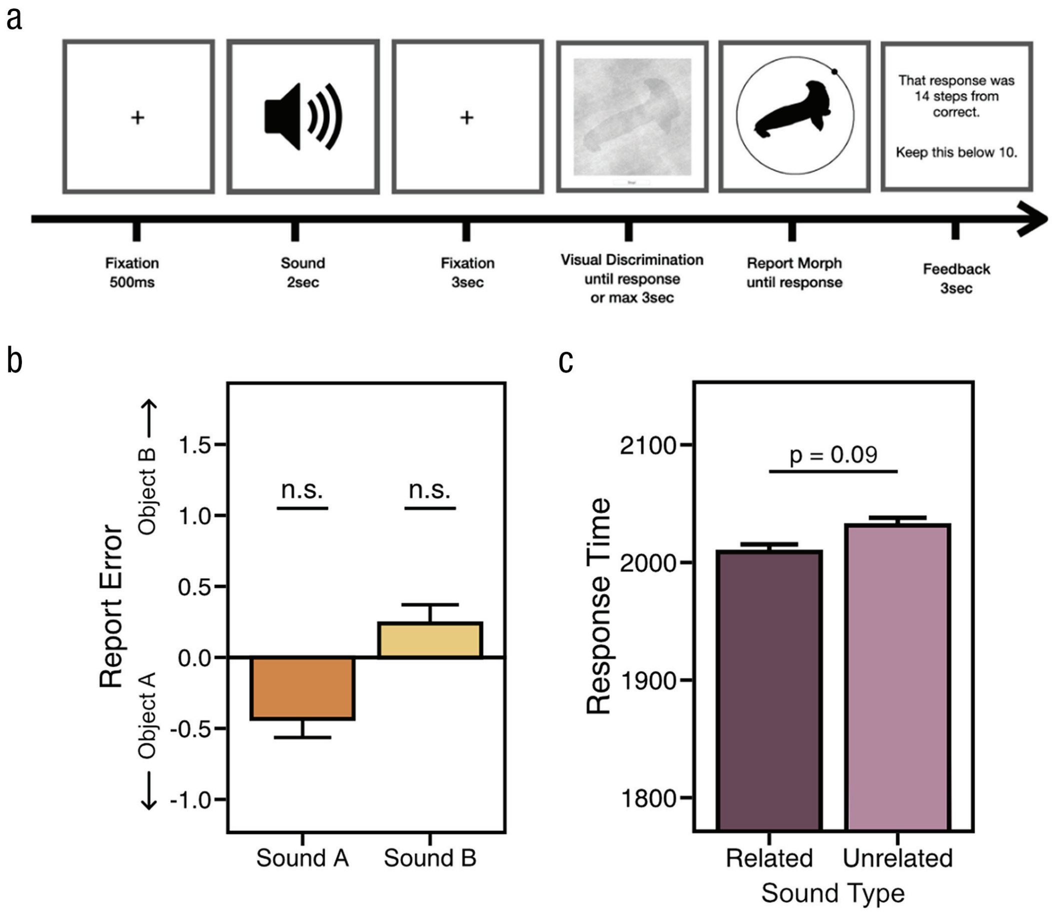


Fig. 4. Results from Experiment 3. (a) Sounds were played prior to the onset of the visual discrimination phase. (b) Report error. We found a substantially reduced, nonsignificant effect (Cohen's d_z = 0.13) for sound A and sound B, suggesting that expectation and attention did not play a substantial role in the effects observed in Experiments 1a, 1b, and 2b. (c) Response times for related and unrelated conditions. The error bars for all data figures are ±1 SEM.

Discussion

Our results suggest that naturalistic auditory information alters the representations of objects we see. Specifically, we found that visual features of object representations are shifted toward features that are congruent with a concurrent auditory stimulus: The same ambiguous object (e.g., a 50% seal and 50% hammer morph) was perceived as more hammer-like when paired with a hammer sound and more seal-like when paired with the sound of seal barking. In a series of control experiments, we demonstrated that these cross-modal effects are not due to biases at decision nor response stages (Experiments 2a and 2b), nor is explicit semantic knowledge about the sounds sufficient to elicit these effects (e.g., volitional search for specific features; Experiment 3). Instead, sounds exert a reliable effect on visual perception only when both stimuli overlap temporally. Additionally, and broadly consistent with other research on this topic, our findings revealed that the sounds hasten the accumulation of related visual information, resulting in faster RTs for related, relative to unrelated, audiovisual inputs.

How might sounds exert influences over visual perception? In the natural world, sounds are causally predictive of the object that generated them—cats cannot bark, for example—and thus, sounds provide independent and informative cues about the visual world. This reliable and highly predictive relationship between audiovisual events can drive changes in early visual processing regions of the brain (van Atteveldt et al., 2014), leading to selective processing of relevant visual features. Previous work has shown that auditory information can rapidly affect the earliest stages of visual processing (Giard & Peronnet, 1999), that auditory and visual signals are integrated in a near-optimal way (Alais & Burr, 2004; Aller & Noppeney, 2019; Burr et al., 2009), that predictive relationships between stimuli lead to a selective reweighting of probabilistically relevant features (Bell et al., 2016; de Lange et al., 2018; Kok et al., 2012), and that these effects are largely driven by previous experience (Gau & Noppeney, 2016; Seriès & Seitz, 2013; Stocker & Simoncelli, 2006). For example, Kok and colleagues (2012) showed that when sensory information predicts an event, processing of probabilistically irrelevant features is suppressed relative to relevant features—those that are more likely to be observed—ultimately sharpening the processing of relevant sensory information. Taken together, these results led us to hypothesize that the clear sounds presented in our study exerted a strong influence over early visual processing, which led to a selective modulation of visual features that were inferred to come from the same generative object (i.e., ambiguous features are presumed to be dog-like when co-occurring with the sound of a barking dog). This suggests that naturalistic sounds do not simply hasten visual perception but that this speed decrease may be the result of shifting perceptual representations toward expected visual features. Additionally, within this framework, such sharpening of sensory processing can also lead to a facilitation of visual feature extraction for expected features, as evidenced by faster RTs for related relative to unrelated sounds.

Another possible source of this effect may be that high-level semantic knowledge influences visual perception (Chen & Spence, 2011a). For example, presenting linguistic labels prior to a visual object has been shown to boost perceptual processing (Lupyan & Ward, 2013). However, the present results are inconsistent with the hypothesis that activating semantic knowledge underlies the perceptual changes we observed here, because the semantic content of real-world sounds alone did not reliably shift perceptual representations (Experiment 3). Our results support the more implicit and low-level process of probabilistic inference (Seriès & Seitz, 2013), where the purported effects of semantics and top-down goals on visual perception operate through separate mechanisms (Cox & Hong, 2015; Gordon et al., 2019; Helbig & Ernst, 2008). Furthermore, finding that audiovisual events need to overlap temporally to exert an effect is also in line with the notion that the learned structure from the world—here, that sounds are exclusively produced by appropriate objects and that audiovisual events co-occur in time—influences how we perceive novel sensory information (Summerfield & Egner, 2009).

Our results broadly relate to work that has shown influences of auditory context on visual-perceptual processing for realistic objects. However, in previous work, the crossmodal facilitation of visual perception (a) was often observed after explicit familiarization or training with the audiovisual stimuli; (b) was often observed with a task that required participants to report whether the sound and image were congruent, thus examining RT and accuracy rather than perceptual biases; and (c) typically involved rapid presentation of the visual stimulus—where some trials might represent uncertain or low-confidence perception, possibly resulting in biases or specific response strategies (Chen & Spence, 2011a, 2018b; Schneider et al., 2008). Here, we avoided these potential limitations and designed a task with a unique stimulus set that allowed us to measure more naturally occurring crossmodal effects and assess the perceptual representations themselves. In particular, (a) participants received no training and had no direct experience with the experimental stimuli prior to participating; (b) the task entailed and encouraged participants to accurately report the visual target irrespective of the audiovisual relationship, thus avoiding potential congruency biases; and (c) participants were in control of the amount of visual information they accumulated, thus allowing us to more confidently assume that participants had sufficient visual information to complete each trial accurately. Importantly, this last point demonstrates that this crossmodal effect is not limited to especially noisy perceptual representations, nor are they the product of uncertainty at response (especially because participants were encouraged on every trial to keep their error as low as possible), suggesting that well-formed perceptual representations are nonetheless influenced by auditory context.

Overall, our findings demonstrate that the ongoing perceptual processing of novel and ambiguous stimuli is altered by related auditory context such that the ultimate perceptual representation is shifted toward sound-congruent features. Our results favor a multisensory rather than a decisional or strategic account, in which visual and auditory information are continuously integrated such that inputs from one modality—in our case audition—trigger inferences about the world that the visual system uses to interpret concurrent ambiguous information. Most broadly, our study demonstrates the importance of investigating visual processing as an integrative rather than an isolated process (Körding et al., 2007) and that multisensory integration plays a critical role in forming visual object representations.

Transparency

Action Editor: Rachael Jack
Editor: Patricia J. Bauer
Author Contributions

J. R. Williams: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, validation, visualization, writing – original draft, writing – review and editing.

Y. A. Markov: conceptualization, software, writing – review and editing.

N. A. Tiurina: conceptualization, software, writing – review and editing.

V. S. Störmer: conceptualization, funding acquisition, methodology, project administration, supervision, validation, writing – review and editing.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This work was supported by the National Science Foundation Graduate Research Fellowship Program (Grant No. DGE-2038238).

Open Practices

All data, scripts, and stimuli have been made publicly available via OSF and can be accessed at https://osf.io/85kwv. The design and analysis plans for the experiments were not preregistered. This article has received the badges for Open Data and Open Materials. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

ORCID iD

Jamal R. Williams https://orcid.org/0000-0002-3034-511X

Acknowledgments

We thank Timothy Brady, Edward Vul, Maria Robinson, and Jonathan Keefe for their helpful conversations, comments, and suggestions.

Note

[1] Note that the anchor objects were never targets, and the visual and auditory stimuli were presented concurrently to capitalize on the tight temporal integration window during multisensory integration (Chen & Spence, 2018a; Edmiston & Lupyan, 2015).

References

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology*, *14*(3), 257–262. https://doi.org/10.1016/j.cub.2004.01.029

Aller, M., & Noppeney, U. (2019). To integrate or not to integrate: Temporal dynamics of hierarchical Bayesian causal inference. *PLOS Biology*, *17*(4), Article e3000210. https://doi.org/10.1371/journal.pbio.3000210

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience*, *5*(8), 617–629. https://doi.org/10.1038/nrn1476

Bell, A. H., Summerfield, C., Morin, E. L., Malecek, N. J., & Ungerleider, L. G. (2016). Encoding of stimulus probability in macaque inferior temporal cortex. *Current Biology*, *26*(17), 2280–2290. https://doi.org/10.1016/j.cub.2016.07.007

Biederman, I., Glass, A. L., & Stacy, E. W. (1973). Searching for objects in real-world scenes. *Journal of Experimental Psychology*, *97*(1), 22–27. https://doi.org/10.1037/h0033776

Boutonnet, B., & Lupyan, G. (2015). Words jump-start vision: A label advantage in object recognition. *Journal of Neuroscience*, *35*(25), 9329–9335. https://doi.org/10.1523/JNEUROSCI.5111-14.2015

Burr, D., Banks, M. S., & Morrone, M. C. (2009). Auditory dominance over vision in the perception of interval duration. *Experimental Brain Research*, *198*(1), 49–57. https://doi.org/10.1007/s00221-009-1933-z

Champely, S., Ekstrom, C., Dalgaard, P., Gill, J., Weibelzahl, S., Anandkumar, A., Ford, C., Volcic, R., De Rosario, H. and De Rosario, M. H. (2018). Package 'pwr'. *R package version*, *1*(2).

Chen, Y. C., & Spence, C. (2010). When hearing the bark helps to identify the dog: Semantically-congruent sounds modulate the identification of masked pictures. *Cognition*, *114*(3), 389–404. https://doi.org/10.1016/j.cognition.2009.10.012

Chen, Y. C., & Spence, C. (2011a). The crossmodal facilitation of visual object representations by sound: Evidence from the backward masking paradigm. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(6), 1784–1802. https://doi.org/10.1037/a0025638

Chen, Y. C., & Spence, C. (2011b). Crossmodal semantic priming by naturalistic sounds and spoken words enhances visual sensitivity. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(5), 1554–1568. https://doi.org/10.1037/a0024329

Chen, Y. C., & Spence, C. (2018a). Audiovisual semantic interactions between linguistic and nonlinguistic stimuli: The time-courses and categorical specificity. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(10), 1488–1507. https://doi.org/10.1037/xhp0000545

Chen, Y. C., & Spence, C. (2018b). Dissociating the time courses of the cross-modal semantic priming effects elicited by naturalistic sounds and spoken words. *Psychonomic Bulletin and Review*, *25*(3), 1138–1146. https://doi.org/10.3758/s13423-017-1324-6

Colonius, H., & Diederich, A. (2004). Multisensory interaction in saccadic reaction time: A time-window-of-integration model. *Journal of Cognitive Neuroscience*, *16*(6), 1000–1009. https://doi.org/10.1162/0898929041502733

Cox, D., & Hong, S. W. (2015). Semantic-based crossmodal processing during visual suppression. *Frontiers in Psychology*, 6, Article 722. https://doi.org/10.3389/fpsyg.2015.00722

Davenport, J. L., & Potter, M. C. (2004). Scene consistency in object and background perception. *Psychological Science*, *15*(8), 559–564.

de Lange, F. P., Heilbron, M., & Kok, P. (2018). How do expectations shape perception? *Trends in Cognitive Sciences*, *22*(9), 764–779. https://doi.org/10.1016/j.tics.2018.06.002

Edmiston, P., & Lupyan, G. (2015). What makes words special? Words as unmotivated cues. *Cognition*, 143, 93–100. https://doi.org/10.1016/j.cognition.2015.06.008

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429–433. https://doi.org/10.1038/415429a

Gau, R., & Noppeney, U. (2016). How prior expectations shape multisensory perception. *NeuroImage*, 124, 876–886. https://doi.org/10.1016/j.neuroimage.2015.09.045

Giard, M.-H., & Peronnet, F. (1999). Auditory-visual integration during multimodal object recognition in humans: A behavioral and electrophysiological study. *Journal of Cognitive Neuroscience*, *11*(5), 473–490.

Gordon, N., Tsuchiya, N., Koenig-Robert, R., & Hohwy, J. (2019). Expectation and attention increase the integration of top-down and bottom-up signals in perception through different pathways. *PLOS Biology*, *17*(4), Article e3000233. https://doi.org/10.1371/journal.pbio.3000233

Helbig, H. B., & Ernst, M. O. (2008). Visual-haptic cue weighting is independent of modality-specific attention. *Journal of Vision*, *8*(1), Article 21. https://doi.org/10.1167/8.1.21

Heron, J., Whitaker, D., & McGraw, P. V. (2004). Sensory uncertainty governs the extent of audio-visual interaction. *Vision Research*, *44*(25), 2875–2884. https://doi.org/10.1016/j.visres.2004.07.001

Kok, P., Jehee, J. F. M., & de Lange, F. P. (2012). Less is more: Expectation sharpens representations in the primary visual cortex. *Neuron*, *75*(2), 265–270. https://doi.org/10.1016/j.neuron.2012.04.034

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal inference in multisensory perception. *PLOS ONE*, *2*(9), Article e943. https://doi.org/10.1371/journal.pone.0000943

Liao, J., Lima, R. S., Nehab, D., Hoppe, H., Sander, P. V., & Yu, J. (2014). Automating image morphing using structural similarity on a halfway domain. *ACM Transactions on Graphics*, *33*(5), Article 168. https://doi.org/10.1145/2629494

Lupyan, G., & Ward, E. J. (2013). Language can boost otherwise unseen objects into visual awareness. *Proceedings of the National Academy of Sciences, USA*, *110*(35), 14196–14201. https://doi.org/10.1073/pnas.1303312110

McDonald, J. J., Teder-Saälejärvi, W. A., & Hillyard, S. A. (2000). Involuntary orienting to sound improves visual perception. *Nature*, *407*(6806), 906–908. https://doi.org/10.1038/35038085

Meredith, M. A., Nemitz, J. W., & & Stein, B. E. (1987). Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors. *Journal of Neuroscience*, *7*(10), 3215–3229. https://doi.org/10.1523/jneurosci.07-10-03215.1987

Rohe, T., & Noppeney, U. (2015). Sensory reliability shapes perceptual inference via two mechanisms. *Journal of Vision*, *15*(5), Article 22. https://doi.org/10.1167/15.5.22

Sadr, J., & Sinha, P. (2004). Object recognition and random image structure evolution. *Cognitive Science*, *28*(2), 259–287. https://doi.org/10.1016/j.cogsci.2003.09.003

Schneider, T. R., Engel, A. K., & Debener, S. (2008). Multisensory identification of natural objects in a two-way crossmodal priming paradigm. *Experimental Psychology*, *55*(2), 121–132. https://doi.org/10.1027/1618-3169.55.2.121

Sekuler, R., Sekular, A. B., & Lau, R. (1997). Sound alters visual motor perception. *Nature*, *385*(23), 308.

Seriès, P., & Seitz, A. R. (2013). Learning what to expect (in visual perception). *Frontiers in Human Neuroscience*, *7*, Article 668. https://doi.org/10.3389/fnhum.2013.00668

Stocker, A. A., & Simoncelli, E. P. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience*, *9*(4), 578–585. https://doi.org/10.1038/nn1669

Störmer, V. S., McDonald, J. J., & Hillyard, S. A. (2009). Cross-modal cueing of attention alters appearance and early cortical processing of visual stimuli. *Proceedings of the National Academy of Sciences, USA*, *106*(52), 22456–22461. https://doi.org/10.1073/pnas.0907573106

Summerfield, C., & Egner, T. (2009). Expectation (and attention) in visual cognition. *Trends in Cognitive Sciences*, *13*(9), 403–409. https://doi.org/10.1016/j.tics.2009.06.003

van Atteveldt, N., Formisano, E., Goebel, R., & Blomert, L. (2007). Top-down task effects overrule automatic multisensory responses to letter-sound pairs in auditory association cortex. *NeuroImage*, *36*(4), 1345–1360. https://doi.org/10.1016/j.neuroimage.2007.03.065

van Atteveldt, N., Murray, M. M., Thut, G., & Schroeder, C. E. (2014). Multisensory integration: Flexible use of general operations. *Neuron*, *81*(6), 1240–1253. https://doi.org/10.1016/j.neuron.2014.02.044

Vroomen, J., & De Gelder, B. (2000). Sound enhances visual perception: Cross-modal effects of auditory organization on vision. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(5), 1583–1590. https://doi.org/10.1037/0096-1523.26.5.1583

Watanabe, K., & Shimojo, S. (2001). When sound affects vision: Effects of auditory grouping on visual motion perception. *Psychological Science*, *12*(2), 109–116. https://doi.org/10.1111/1467-9280.00319

Williams, J. R., & Störmer, V. S. (2019). Auditory information facilitates sensory evidence accumulation during visual object recognition. *Journal of Vision*, *19*(10), Article 20c. https://doi.org/10.1167/19.10.20c

Acknowledgements

Chapter 3, in full, is a reprint of the material as it appears in Psychological Science, 2022, Williams, Jamal R.; Markov, Yuri A.; Tiurina, Natalia A.; Störmer, Viola S. The dissertation author was the primary investigator and author of this paper.

CONCLUSION

Given the complexity of achieving visual perception, Zeno's dichotomy seems applicable[3]: since the process of converting photons to perceptions requires a seemingly infinite number of intermediate steps perception must be an illusion. Paradoxes notwithstanding, visual perception *feels* as though it occurs instantaneously and with little effort in spite of the many steps required to produce it. To facilitate processing and thereby alleviate the immense burden of these computations, visual processing incorporates prior knowledge and available context. Across three chapters, I demonstrate that priors and context, influence our ability to process the visual world. In the real world, context likely facilitates our ability to infer object identity (Oliva & Torralba, 2007), the configuration of occluded objects (Sekuler & Palmer, 1992), and, among many other things, whether a dress is white and gold or blue and black (Lafer-Sousa et al., 2015). Yet, these attempts to resolve visual ambiguity can lead to an unavoidable capture of attention, a beneficial facilitation of object recognition, and a perturbation of object recognition.

Broadly, chapter 1 demonstrates that information active in working memory can capture and guide attention towards matching features in the environment. Suggesting a critical role of internally activated context in attentional selection. Beyond the effects of visual content affecting visual processing, in chapter 2, direct and indirect auditory information increased the rate by which meaningful visual information was acquired. However, chapter 3 suggests that this facilitation is not cost free. Instead, we show that—even when people self-report as having a complete, and precise perceptual representation—this facilitatory process is likely driven by predictions about impending visual information that are capable of biasing perception.

[3] I swear this is the last time I'll bring up the ancient Greeks.

The integration of these findings at large undoubtedly informs our understanding of how context, whether within or across modalities, affects how we process and perceive visual information. For example, in chapter 1, we set out to answer several outstanding questions in the visual search literature. Historically, the discussion of whether multiple working memory items can guide attention has often focused on the number of items that can achieve a privileged template status with little focus on how well these items are remembered. In this chapter, in contrast, a careful assessment of the memory strength of remembered items, demonstrated a rather straightforward relationship between memory strength (the average representational fidelity of memories) and attentional guidance. This direct relationship exists independently of whether an item's status is elevated to a "template". In particular, this work showed that both within- and across-trial variation in how well items are represented predicted the strength of attentional guidance and did so without needing to posit any other predictors (like a special privileged state). These findings suggest that the degree to which an item accurately represents the originally encoded item (i.e., its representational fidelity) determines whether—and how effectively—an item guides attention.

The "representational fidelity" framework discussed in chapter 1 speaks to two important issues in the literature of memory driven attentional guidance as well as to working memory and attention literature more broadly. First, to the question of whether one, or many working memory representations guide attention. Our data indicate that only an extremely strong and high-fidelity memory representation can guide visual search effectively, something that might rarely occur for more than one item at a time. To be clear, it could be the case that multiple items simultaneously guide attention, however, the data on offer simply suggests that all of the simultaneously maintained memories are unlikely to be sufficiently well-represented to each exert strong and reliably measured guidance over attention in these sorts of tasks (*see future directions*). Second, the present results elucidate the mechanisms underlying attentional guidance and explain why attentional guidance is often driven by a single item. Importantly, and different from previous accounts, this data suggests that natural variation in the representational fidelity between items is sufficient to explain the extent to which an item will guide attention at any given time, with no special focus of attention or similar state-based accounts of working memory being necessary. In sum, these findings have important implications for our understanding of the fundamental structure and processes involved in working memory and attention.

Chapter 2 explored whether naturalistic sounds affected visual object processing by accelerating how quickly relevant visual features are extracted from noisy visual input. We used a visual-discrimination task and a perceptual test of object recognition to demonstrate that auditory information can hasten visual-feature extraction without negatively impacting accuracy. With this paradigm, we show that participants responded more quickly in the visual-discrimination phase when a non-predictive sound was congruent with the target image compared to when it was incongruent, suggesting that participants demanded a greater amount of visual information to perform the 2AFC test for incongruent compared to congruent trials. These results reveal that congruent auditory object and scene information facilitate visual object processing when compared to incongruent sounds, and they further demonstrate that these cross-modal effects occur for both categorical and detailed object recognition. This latter finding suggests that naturalistic sounds are not simply supporting broad visual categorization (i.e., is it a bird or not?) but can also help to extract detailed visual-feature information as well (i.e., is it this bird or that one?). Taken together, these experiments reveal that visual perception is not only affected by the direct relationship between visual objects and the sounds that they make but that more abstract auditory information can be leveraged to enhance the extraction of meaningful and detailed visual features.

Overall, chapter 2 demonstrates that incidental (ie, task-irrelevant) auditory objects and scenes facilitates visual perception of related visual objects, even when sounds are not predictive of the visual stimulus in a given task and even when participants have no prior experience with the particular set of audiovisual stimuli. Collectively, our results suggest that perception integrates contextual information at various levels of processing and can leverage general, gist-like information across sensory modalities to facilitate visual object perception.

The results from chapter 2 suggest that when sights and sounds align, visual perception is enhanced. Given the nature of the task, I concluded that feature extraction was performed more quickly on congruent compared to incongruent trials. But the hypothesis here is that this facilitation might be driven by an implicit prediction about the congruency between object sounds and the object that is capable of producing them (eg, it is natural to expect ducks, not lions, when we hear a series of quacks). Indeed, chapter 3 suggests that naturalistic auditory information alters the representations of objects we see. Specifically, visual features of object representations are shifted toward features that are congruent with a concurrent auditory stimulus: The same ambiguous object (e.g., a 50% seal and 50% hammer morph) was perceived as more hammer-like when paired with a hammer sound and more seal-like when paired with the sound of seal barking. In a series of control experiments, this perturbation of a complete and confident perceptual representation was unlikely to be driven by biases at decisional/response stages nor were they likely driven by top-down, volitional attentional selection. That is, it is unlikely to be driven in any large part by pre- or post-perceptual processes and is instead a mechanism that happens during the processing of novel visual information as sounds only exert a reliable effect on visual perception when both stimuli overlap temporally, as a unified audiovisual object.

Overall, chapter 3 demonstrates that the ongoing perceptual processing of novel and ambiguous stimuli is altered by related auditory context such that the ultimate perceptual representation is shifted toward sound-congruent features. Our results favor a multisensory rather than a decisional or strategic account, in which visual and auditory information are continuously integrated such that inputs from one modality—in our case audition—trigger inferences about the world that the visual system uses to interpret concurrent ambiguous information. Most broadly, this chapter demonstrates the importance of investigating visual processing as an integrative rather than an isolated process (Körding et al., 2007) and that multisensory integration plays a critical role in forming visual object representations.

Context and the quality of internal and external information

Thus far I have asserted that, despite our intuition, the world is full of ambiguous and indeterminate sensory information that must be resolved to effectively perceive our environment. To resolve this ambiguity, the quality of the internal and external information is weighed and integrated according to its relevance (Friston, 2005; Kersten et al., 2004; De Lange et al., 2018). Across these chapters I have suggested that the quality of the internal (ch 1) and external (ch2-3) context defines the strength and reliability of this integration process. Specifically, in chapter 1, the quality of the internal representation is shown to directly relate to how well it interacts with and captures attention. In the following chapters, a clear and unambiguous source of external context is provided to ensure that this information is integrated as the noisy visual input is resolved internally. However, while many questions remain unanswered, a critical component to pursue in this body of work is to better characterize how the independent quality of internal and external information interact with one another and influence visual processing.

As a general framework (this and other work alludes to the premise that) the quality or reliability of primary, target information plays a central role in determining the extent to which secondary information is capable of affecting visual processing (Alais & Burr, 2004; Noppeney, 2021; Shams & Kim, 2010). In chapter 1, the secondary, internal information was modulated while the external, primary information remained reliable and precise. Yet, within the visual search literature, it is unclear how a strong, well-represented memory will interact with items in the environment if we modulate how well they match to the internal representation or modulate their quality and reliability. Some theories suggest that as the internal representation becomes more precise, its ability to interact with weak, secondary information in the environment decreases (*see* Yu et al., 2023) while others posit the exact opposite effect (Williams et al., 2023; *more in future directions*). In chapters 2-3, the opposite pattern was explored: primary information was degraded while secondary information remained clear and unambiguous. When primary information is degraded (as in ch. 2-3), reliable auditory information was thoroughly integrated to resolve the lack of clarity. Yet it remains unclear how information might be integrated when both internal and external sources share roughly equal quality and reliability? Or when they are independently modulated?

In chapter 2, people were given control over the amount of visual information they received. This provides a unique insight into how secondary information is integrated as the primary information becomes clearer and more identifiable. Since, it is presumed that a relatively strong, and reliable perception was achieved, I conclude that the auditory information helped to speed the resolution process but can say little about the perceptual representation other than they were sufficiently strong to complete the final test. However, while chapter 1 showed that high fidelity is critical to allow for interactions with attention, interactions with perception might not have this requirement: Even when participants were given control over how well the target item was presented, in chapter 3, we nevertheless saw reliable biases in their representations. Thus, the interactions between contextual influence and visual perception, attention, and fidelity of the internal representation deserve far more investigations.

Future directions

As of yet, it remains unclear how multiple working memory items might guide attention at the same time. This question has been given limited priority in the literature and has not been examined with careful measurement of representational fidelity and memory strength (Hollingworth & Beck, 2016). In preliminary studies, I have found that both memories interact with attention in a seemingly additive way: guidance when two items are present is twice as strong compared to when only a single memory item is present. Similar to Experiment 3 in Chapter 1, my next goal is to modulate the fidelity of both items equally and map the effects of strength onto guidance for the full contents of visual working memory. This, as more thoroughly described in Chapter 1, is in stark contrast to the common theory in the field that guidance is determined by a special status within working memory.

Since, even the inclusion of, two items can complicate the outcome of the, currently, simple modeling process (*described in chapter 1*), when a single item is maintained in working memory, we can more assuredly modulate the strength of this item. Since we can test a single item at the same time as when a search trial would have appeared, we do not need to make presumptions about the average quality of the item at the time of search—as we often need to do with secondary items—and instead can modulate and measure the strength of the item in question. As we modulate this items memory strength, we can map the population level activity that represents a given feature and can map how the shape of this activity changes with strength.

While fidelity seems critical for the strength and reliability of a memory's interactions with attention, fidelity appears to primarily affect the strength of the cross modal contextual effect over perception. In explorations of reliability, when control was taken from participants and modulated experimentally, we found that the strength of the perceptual perturbation effect would increase with strength as visual information decreased and decrease as information increased, but never diminished to zero. Suggesting that concurrently available information can bias perception even when fidelity is high (or at least sufficient to do the continuous report task). Additionally, we are exploring how visual perceptions develop neurally using electroencephalogram (EEG). In this work, we modulate how rapidly visual information is uncovered from noise while participants perform the same discrimination task as before. Our initial results suggest that meaningful information is accumulated more quickly when congruent object information accompanies a visual target, in line with our behavioral intuitions. However, as in our other cross modal work, visual information starts in a degraded state and is stopped per individual when the visual information is sufficient. In the future, we plan to manually control the level of visual information that has accumulated per trial and map the psychometric functions along with corresponding ERP components. This will help us to demonstrate how context affects visual processing when the fidelity of one sensory stream is held constant and the other is experimentally modulated.

Concluding remarks

Across three chapters I demonstrate that visual processing can leverage the context from multiple senses to guide, facilitate, and perturb perception and attention; likely in an effort to lessen the intense burden required to process the massive amount of visual information at any given moment. These findings are indeed limited in scope and further work should explore several facets. In particular: list things here. That being well-said it is important to consider how visual perception is modulated and augmented by our experiences in the real world.

REFERENCES

Alais, D., & Burr, D. (2004). The Ventriloquist Effect Results from Near-Optimal Bimodal Integration. Current Biology, 14(3), 257–262. https://doi.org/10.1016/j.cub.2004.01.029

Burr, D., Banks, M. S., & Morrone, M. C. (2009). Auditory dominance over vision in the perception of interval duration. Experimental Brain Research, 198(1), 49–57. https://doi.org/10.1007/s00221-009-1933-z

De Lange, F. P., Heilbron, M., & Kok, P. (2018). How do expectations shape perception?. *Trends in cognitive sciences*, *22*(9), 764-779.

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1456), 815–836. https://doi.org/10.1098/rstb.2005.1622

Hollingworth, A., & Beck, V. M. (2016). Memory-based attention capture when multiple items are maintained in visual working memory. Journal of Experimental Psychology: Human Perception and Performance, 42(7), 911–917. https://doi.org/10.1037/xhp0000230

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object Perception as Bayesian Inference. Annual Review of Psychology, 55(1), 271–304. https://doi.org/10.1146/annurev.psych.55.090902.142005

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal Inference in Multisensory Perception. PLoS ONE, 2(9), e943. https://doi.org/10.1371/journal.pone.0000943

Lafer-Sousa, R., Hermann, K. L., & Conway, B. R. (2015). Striking individual differences in color perception uncovered by 'the dress' photograph. *Current Biology*, *25*(13), R545-R546.

Noppeney, U. (2021). Perceptual Inference, Learning, and Attention in a Multisensory World. *Annual Review of Neuroscience*, *44*(1), 449–473. https://doi.org/10.1146/annurev-neuro-100120-085519

Oliva, A., & Torralba, A. (2007). The role of context in object recognition. Trends in Cognitive Sciences, 11(12), 520–527. https://doi.org/10.1016/j.tics.2007.09.009

Shams, L., & Kim, R. (2010). Crossmodal influences on visual perception. *Physics of Life Reviews*, *7*(3), 269–284. https://doi.org/10.1016/j.plrev.2010.04.006

Sekuler, A. B., & Palmer, S. E. (1992). Perception of partly occluded objects: A microgenetic analysis. Journal of Experimental Psychology: General, 121(1), 95–111. https://doi.org/10.1037/0096-3445.121.1.95

Williams, J., Brady, T., & Stoermer, V. (2023). Precise Memories and Imprecise Guidance: Why attention is guided towards colors that I'm certain I didn't see. *Journal of Vision*, *23*(9), 5957-5957.

Yu, X., Zhou, Z., Becker, S. I., Boettcher, S. E. P., & Geng, J. J. (2023). Good-enough attentional guidance. Trends in Cognitive Sciences, 27(4), 391–403. https://doi.org/10.1016/j.tics.2023.01.007