UNIVERSITY OF CALIFORNIA SAN DIEGO

Learning and Prior Knowledge Shape Cognitive Representations for Complex Images

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

in

Experimental Psychology

by

Hayden Schill

Committee in charge:

Professor Timothy Brady, Chair
Professor Anastasia Kiyonaga
Professor Viola Stoermer
Professor John Wixted

2023

The dissertation of Hayden Schill is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

DEDICATION

To my family and mentors

EPIGRAPH

It is a well known experience that one and the same object seldom occasions precisely similar perceptions in the minds of different people.

Of the same landscape the poet's image would differ greatly from that of the botanist, the painter's from that of the geologist or the farmer, the stranger's from that of him who calls it home. In the same way, one and the same speech is often understood in as many different ways as there are hearers. What does not the child see in his toys, the devout mind in the objects of its devotions! What does not the experienced reader of human nature see in the wrinkles and folds, the wilted and weather-beaten features, of a human face! How much do the gestures, the play of features, the glowing or fading fire of the eye, tell him of the battles and storms of the soul! And the artist, does he not perceive in a work of art a thousand things that escape the closest attention of the ordinary observer? Has not each of us the sharpest kind of an eye for the objects with which our calling makes us best acquainted?

There are as many different ideas of one and the same thing as there are observers. Whence this variation in apprehension, with otherwise similar sense apparatus?

Carl Lange, 1894

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOLEDGEMENTS

I would first like to thank Professor Tim Brady, my Ph.D. advisor and chair of my dissertation committee. Tim is rightfully known not only as a pioneering scientist, but also as a wonderful mentor, collaborator, and instructor. I am greatly appreciative of Tim's support not only of my academic growth, but of my professional and leadership endeavors as well. His positivity and flexibility were indispensable through the COVID-19 pandemic. I am proud to have had the opportunity to learn from Tim's example and feel fortunate to have such a wonderful collaborator and mentor.

I would also like to thank the other members of my dissertation committee – Dr. Anastasia Kiyonaga and Dr. Viola Störmer, the first women faculty mentors I've had the opportunity to work with, and Dr. John Wixted, whose passion for justice and theory is contagious – I can't wait to teach Psychology and the Law next spring at UC Riverside. The engaging conversations with my committee over the last few years have inspired my curiosity and helped me grow as researcher.

My graduate school experience has been marked with a sense of community and friendship, and for that I would like to thank my lab mates from the Vision & Memory Lab, and in particular Isabella Destefano and Jamal Williams, who have been through every step of the program with me. Thanks to Maria Robinson, Malinda McPherson, Kirsten Adams, post-docs who served as great examples of strong scientists and mentors. Thanks to many others who have been friends, support networks, and builders of community – Michael Allen, Chaipat Chunharas, Jonas Lau, Lauren Williams, Mark Schurgin, Anna Shafer-Skelton, Angus Chapman, Jonathan Keefe, Janna Wennberg, and more.

I would also like to thank my research assistants in chronological order – Samantha Gray, Ani Abovian, and Natalia Pallis-Hassani. Sam Gray and I collaborated on several projects together, including Chapter 2 which investigates visual hindsight bias in expert radiologists. Along with an incredible breadth of interests and the skillset to accomplish projects ranging from medical image perception to intracranial recordings, Sam is one of the most engaged and organized collaborators I have worked with. Ani, now a PhD student at UC Davis, excelled in crafting well thought out experimental designs that were well rooted in the literature. Natalia has been my most recent RA collaborator, and I am inspired by her eagerness and aptitude to learn new skills and have enjoyed watching her scientific journey unfold. It is with these remarkable researcher's help that this dissertation is made possible.

I would also like to thank those who supported me on my path to graduate school, especially Dr. Jeremy Wolfe and Dr. Jason Haberman. Joining the Wolfelab inspired my interest in bringing the methods and questions of visual cognition into applied domains such as radiology. Thank you to Jason for taking a chance on a 3rd year undergraduate student who did not know what research was, for inspiring my interest in ensemble perception, and for encouragement and advice I pass down to my own students I work with. I am eternally grateful to have lifelong mentors such as these.

Thank you to my family for their support of my education and endeavors over the years – my grandparents, parents, and my siblings, especially to my mom Tammy and sister Taylor, without their help over the last year this dissertation might not have come together. Finally, thank you to my husband Sam, who has been here throughout it all, even from afar.

A few last groups I'd like to thank: First, the students in my classes, especially Applied Cognitive Psychology, for taking a chance on a new course and new instructor.

Their participation and discussion taught me so much about being an instructor and allowed me to recognize my own love of teaching. Thanks to funding sources, including the National Science Foundation and the P.E.O. Scholar Award. I'd like to especially thank Chapter K for their support over this last year. Finally, thank you to all my advocacy partners, and in particular Patriccia Ordoñez-Kim and Ernesto Arciniega. Advocating alongside many passionate students and staff and growing in partnerships has been one of the most rewarding experiences over the last few years, and I am thankful for the opportunities it has given me to learn and grow as a leader.

Chapter 1, in full, is a reprint of the material as it appears in Memory and Cognition, 2021, co-authored with Dr. Jeremy M. Wolfe and Dr. Timothy F. Brady. The dissertation author was the primary author of this paper.

Chapter 2, in full, is a reprint of the material as it appears in the Journal of Medical Imaging, 2023, co-authored with Samantha Gray and Dr. Timothy F. Brady. The dissertation author was the primary author of this paper.

Chapter 3 is co-authored with Natalia Pallis-Hassani and Timothy F. Brady. The dissertation author was the primary author of this chapter.

VITA

2016	Bachelor of Science, Rhodes College
2019-2020	Business Development Specialist, eWallaby
2020	Master of Psychology, University of California San Diego
2022-2023	President & CEO, University of California Graduate & Professional Council
2018-2023	Teaching Assistant, University of California San Diego
2023	Doctor of Philosophy, University of California San Diego

PUBLICATIONS

Schill, H. M., Gray, S. M., & Brady, T. F. (2023). Visual hindsight bias for abnormal mammograms in radiologists. *Journal of Medical Imaging*, *10*(S1). https://doi.org/10.1117/1.JMI.10.S1.S11910

Brady, T. F., Störmer, V. S., Shafer-Skelton, A., Williams, J. R., Chapman, A. F., & Schill, H. M. (2019). Scaling up visual attention and visual working memory to the real world. In *Psychology of Learning and Motivation* (Vol. 70, pp. 29–69). Elsevier. https://doi.org/10.1016/bs.plm.2019.03.001

Schill, H. M., Cain, M. S., Josephs, E. L., & Wolfe, J. M. (2020). Axis of rotation as a basic feature in visual search. *Attention, Perception, & Psychophysics*, *82*(1), 31–43. https://doi.org/10.3758/s13414-019-01834-0

Wolfe, J. M., Alaoui Soce, A., & Schill, H. M. (2017). How did I miss that? Developing mixed hybrid visual search as a ‘model system’ for incidental finding errors in radiology. *Cognitive Research: Principles and Implications*, *2*(1). https://doi.org/10.1186/s41235-017-0072-5

Schill, H. M., Wolfe, J. M., & Brady, T. F. (2021). Relationships between expertise and distinctiveness: Abnormal medical images lead to enhanced memory performance only in experts. *Memory & Cognition*, *49*(6), 1067–1081. https://doi.org/10.3758/s13421-021-01160-7

Schill, H. (2020). Historical Academic Context Gives Key Insights for Graduate Students. Published in the Psychonomic Society Featured Content: Attention, Perception, & Psychophysics Digital Event.

MANUSCRIPTS IN PREPARATION

Schill, H., Pallis-Hassani, N., Brady, T. Ensemble Perception of Faces with Naturalistic Occlusions (in preparation).

FIELD OF STUDY – COLLEGE TEACHING

Associate-in-Lieu (Primary Instructor):

Applied Cognitive Psychology	Summer 2022
Visual Cognition	Summer 2022
Applied Cognitive Psychology	Winter 2023

Guest Speaker:

Introduction to Psychology	Spring 2021
Educational Psychology	Winter 2022

Mini Teaching Activity Symposium Speaker, “Listen-Explain-Write: A Modified Think-Pair-Share for Going Over Exam Questions.” STP Teaching of Psychology Regional Conference. September 2022.

Invited Speaker; UC San Diego Engaged Teaching Hub Summer Teaching Roundtable. “A Scalable Teaching Activity to Enhance Equity and Community in the Classroom.” October 2022.

UC PsychTerms Teaching Roundtable Speaker; “Listen-Explain-Write: Using Evidenced-based Teaching Activities to Promote Equity and Inclusivity in the Classroom.” December 2022.

ABSTRACT OF THE DISSERTATION

Learning and Prior Knowledge Shape Cognitive Representations for Complex Images

by

Hayden Schill

Doctor of Philosophy in Experimental Psychology

University of California San Diego, 2023

Professor Timothy Brady, Chair

Through navigating and interacting with the world, we develop knowledge that can be used to help guide future interactions or process similar information more efficiently. For example, as someone proceeds through medical school and residency to become a radiologist, they develop a deep understanding of medical images. This learned knowledge influences several cognitive processes, which allows them to find abnormalities quickly and efficiently. My dissertation examines factors that may drive how prior knowledge in

perceptual domains influence cognitive processes such as memory and perception. I use a range of stimuli and methods that mimic real-world scenarios and require significant perceptual prior knowledge. In particular, I use medical imaging and face perception as two unique but strong indices of perceptual expertise. In chapter 1, I establish that radiologists have enhanced memory for abnormal compared to normal mammograms. I show evidence that this memory benefit appears to be driven by a unique role of distinctiveness, which emerges with significant prior knowledge in the expert's domain. In chapter 2, I demonstrate that additional knowledge of an abnormality in a medical image not only shapes memory but also enhances experts' perception of medical images themselves, a form of bias called visual hindsight bias. In chapter 3, I examine the impact of occlusions and feature learning on our ability to extract summary statistics of complex facial information. Overall, I highlight how learned knowledge shapes a range of cognitive processes including memory, visual perception, and ensemble perception, and touch on several avenues for future direction.

INTRODUCTION

Whether you are a radiologist deciding whether an x-ray contains an abnormality, or a new professor trying to gauge your classes understanding behind a sea of masks, people encounter complex and incomplete information every day that they must process and make decisions about. The visual system has evolved to efficiently process these complexities and inconsistencies by quickly learning common occurrences and using that information to make inferences, extract summary statistics, or to guide attention when searching a complex scene (e.g., Kundel & La Follette, 1972). How learning influences the processing of new or related information is of interest to a wide array of both applied and basic fields, including medical image perception, educational psychology, and visual search, attention, and memory. This dissertation aims to answer a few outstanding questions related to learning and using perceptual information.

Experts are an ideal population to study prior knowledge's impact on cognitive processes, and the expertise literature has documented several ways in which prior knowledge allows us to maximize the efficiency of limited attentional and visual working memory systems (Curby, Glazek, & Gauthier, 2009; Ericsson & Kintsch, 1995). For instance, becoming an expert in a domain such as chess changes our memory for items in that domain of expertise (Chase & Simon, 1973; de Groot, 1946), allowing us to store more information as long as the information is consistent with the expectations we have formed as a result of our expertise (Gobet & Simon, 1996). Similarly, as medical trainees gain familiarity and expertise in reading medical images, they develop more efficient visual search strategies and use fewer fixations to find the abnormality in the image (Kundel & La Follette, 1972).

There have been several hypotheses proposed for how prior knowledge leads to enhanced cognitive processing for items related to that knowledge. Many authors posit that memory improvement occurs because existing knowledge allows experts to know what variation to expect, through organization of information into a schema, for information in an expert's domain (e.g., Vincente & Want, 1988). Others have suggested that experts in some domains may have developed specialized perceptual processing strategies for objects of their expertise, allowing them to process more information about an item and leader to richer memory traces (Ericsson & Kintsch, 1995; Watson & Robins, 2014; Richler, Wong, & Gauthier, 2011). There is also research that argues that the critical driver of how memorable an item is in each context is its distinctiveness from other items currently being stored in memory (Rawson & Van Overschelde, 2008; Calkins, 1984; Hunt, 2006; Kundel & La Follette, 1972). The way these effects interact has rarely been studied, and many have been studied primarily in domains with limited or no perceptual expertise available (e.g., in word lists).

Chapter 1 and 2 focus on radiologists' memory for mammograms for several reasons: First, search for signs of breast cancer involves a usefully specific perceptual expertise: only 2-3 kinds of local abnormalities are typically present in abnormal mammograms, and radiologists have significant perceptual expertise whether looking at normal or abnormal medical images. This unique prior knowledge is an ideal case study in which to examine the impact of schemas, perceptual processing, and distinctiveness on memory. Additionally, it has been known for decades that many routine errors in radiology are failures of visual perception (Tuddenham & Calvert, 1961; Kundel & Wright, 1969), which contributes to radiology's place at the top of medical specialties sued for negligence

(Baker, 2014). Thus, using well controlled tasks and stimuli in an environment that maintains some of the complexity of real-world scenarios (e.g., seeing how expert perception of mammograms informs decisions) allows a more generalizable and applicable set of knowledge to emerge that speaks towards both theoretical and applied questions (Brady, Störmer, Shafer-Skelton, Williams, Chapman, & Schill, 2019).

Chapter 1 examines memory performance by non-expert novices and expert radiologists for normal versus abnormal mammography images as a case study in understanding the role of schemas, distinctiveness, and expertise in memory. We find that radiologists have enhanced memory for abnormal mammograms, and that this benefit is consistent with a special role of distinctiveness. In other words, the additional knowledge that an expert has compared to a novice gives the expert a boost in memory for abnormal items because they have additional features (i.e., an abnormality), which makes them more distinct in radiologists' memory compared to normal images. This finding is consistent with previous research arguing that the critical driver of how memorable an item is in a given context is its distinctiveness from other items currently being stored in memory (Calkins, 1984; Hunt, 2006).

In chapter 2, I show that additional knowledge of the presence of an abnormality shapes radiologists' perception of medical images themselves, whereby increased knowledge enhances visual perception of known abnormalities compared to abnormal images without additional knowledge about the abnormality. This chapter expands on the results of chapter 1, showing that expertise not only makes the images more distinctive in memory, but also makes the images more distinctive during perception itself. This project has implications for when radiologists are sued for negligence, which is touched on in the chapter as well as in the conclusion. Overall, chapters 1 and 2 compliment and extend research in expertise and visual memory, showing that memory in experts is enhanced by distinctiveness afforded to them with built up perceptual knowledge in their domain.

Chapter 3 examines feature learning, expertise, and noisy information within a new complex stimulus domain: face processing. Face perception is commonly used as a case study in understanding perceptual expertise (Bukach, Gauthier, & Tarr, 2006; Tarr & Gauthier, 2000; Gauthier, Skudlarski, Gore, & Anderson, 2000). Despite being visually similar in pixel space, individual faces look distinctive to us in part because of significant built-up knowledge over time, which stretches their representational space and allows us to notice more minute differences. This learned knowledge results in processing faces holistically, in a similar way as domain experts develop holistic processing of images like mammograms (Kundel, Nodine, Conant, Weinstein, 2007). Indeed, individuals with prosopagnosia, who have lost the ability recognize faces due to damage in face specific areas of the brain, are used to study the loss of expertise (Bukach, Gauthier, & Tarr, 2006; Duchaine et al., 2004).

Face processing, therefore, is a useful way to examine how prior knowledge impacts cognitive processing with even more control over variables than using specialized participant populations such as radiologists. In particular, chapter 3 explores the impact of feature learning and occlusions on our ability to extract summary statistics of faces wearing sunglasses or face masks. While research on face masks have been shown to disrupt holistic processing of individual face perception and lower recognition performance (Freud et al., 2020; Carragher & Hancock, 2020), it is largely unknown how ensemble perception is impacted by incomplete facial information brought upon by more naturalistic occlusions.

Professional expert facial examiners, in a similar way as experts in other domains, are known to employ strategies that emphasize learning which features are most important to distinguishing complex images within their domain (Towler et al., 2021b). Recent research has suggested that individuals may also learn to counteract the disruption of masks in processing individual faces by putting more weight on the available features (Carragher et al., 2022). In addition to asking how naturalistic occlusions impact ensemble perception broadly, chapter 3 also examines the role of feature learning in deriving the ensemble representation for incomplete facial information. The results of chapter 3 not only enhance our understanding of ensemble perception, but also complement the findings of chapter's 1 and 2 by showing perceptual learning impacting a cognitive process in action.

In three chapters, my thesis demonstrates several factors that play a role in how prior knowledge and learning shape our current visual experience. I show how building up knowledge as an expert affords radiologists enhanced memory for abnormal mammograms and found that meaningfulness or distinctiveness drives this benefit in expert's memory. I also provide evidence that experts have a visual hindsight bias when presented with additional information that enhances their perception of medical images. Finally, I also show how adept we are at learning facial features over time, even when faces are partially occluded, and how that influences ensemble perception. This work has both theoretical and applied applications, which are touched on in the conclusion.

References

Baker, S. R. (2014). U.S. medical malpractice: Some data-driven facts. *Springer International Publishing: Notes of a Radiology Watcher*, 177–179.

Brady, T. F., Störmer, V. S., Shafer-Skelton, A., Williams, J. R., Chapman, A. F., & Schill, H. M. (2019). Scaling up visual attention and visual working memory to the real world. In *Psychology of Learning and Motivation* (Vol. 70, pp. 29–69). Elsevier. https://doi.org/10.1016/bs.plm.2019.03.001

Bukach, C. M., Gauthier, I., & Tarr, M. J. (2006). Beyond faces and modularity: The power of an expertise framework. *Trends in Cognitive Sciences*, *10*(4), 159–166. https://doi.org/10.1016/j.tics.2006.02.004

Calkins, M. W. (1894). Experimental. *Psychological Review*, *1*. https://doi.org/10.1037/h0065852

Carragher, D. J., Towler, A., Mileva, V. R., White, D., & Hancock, P. J. B. (2022). Masked face identification is improved by diagnostic feature training. *Cognitive Research: Principles and Implications*, *7*(1), 30. https://doi.org/10.1186/s41235-022-00381-x

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*. https://doi.org/10.1016/0010-0285(73)90004-2

Curby, K. M., Glazek, K., & Gauthier, I. (2009). A visual short-term memory advantage for objects of expertise. *Journal of Experimental Psychology: Human Perception and Performance*, *35*. https://doi.org/10.1037/0096-1523.35.1.94

De Groot, A. D. (1946). Het denken van den schaker: Een experimenteel-psychologische studie *[The thinking of the chess player: An experimental-psychological study]. Noord-Hollandsche Uitgevers Maatschappij.* (n.d.).

Duchaine, B. C., Dingle, K., Butterworth, E., & Nakayama, K. (2004). Normal Greeble Learning in a Severe Case of Developmental Prosopagnosia. *Neuron*, *43*(4), 469–473. https://doi.org/10.1016/j.neuron.2004.08.006

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review*, *102*. https://doi.org/10.1037/0033-295X.102.2.211

Gauthier, I., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000). Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience*, *3*. https://doi.org/10.1038/72140

Gobet, F., & Simon, H. A. (1996). Recall of random and distorted positions: Implications for the theory of expertise. *Memory & Cognition*, *24*. https://doi.org/10.3758/BF03200937

Hunt, R. R. (2006). The concept of distinctiveness in memory research. In R. R. Hunt & J. B. Worthen (Eds.), Distinctiveness and memory (pp. 3–25). *Oxford University Press. Https://doi.org/10.1093/acprof:oso/9780195169669.003.0001*. (n.d.).

Kundel, H. L., & John Wright, D. (1969). The Influence of Prior Knowledge on Visual Search Strategies During the Viewing of Chest Radioqraphs. *Radiology*, *93*(2), 315–320. https://doi.org/10.1148/93.2.315

Kundel, H. L., & La Follette, P. S. (1972). Visual Search Patterns and Experience with Radiological Images. *Radiology*, *103*(3), 523–528. https://doi.org/10.1148/103.3.523

Kundel, H. L., Nodine, C. F., Conant, E. F., & Weinstein, S. P. (2007). Holistic Component of Image Perception in Mammogram Interpretation: Gaze-tracking Study. *Radiology*, *242*(2), 396–402. https://doi.org/10.1148/radiol.2422051997

Rawson, K. A., & Overschelde, J. P. (2008). How does knowledge promote memory? The distinctiveness theory of skilled memory. *Journal of Memory and Language*, *58*. https://doi.org/10.1016/j.jml.2007.08.004

Richler, J. J., Wong, Y. K., & Gauthier, I. (2011). Perceptual expertise as a shift from strategic interference to automatic holistic processing. *Current Directions in Psychological Science*, *20*. https://doi.org/10.1177/0963721411402472

Tarr, M. J., & Gauthier, I. (2000). FFA: A flexible fusiform area for subordinate-level visual processing automatized by expertise. *Nature Neuroscience*, *3*(8), 764–769. https://doi.org/10.1038/77666

Towler, A., Keshwa, M., Ton, B., Kemp, R., & White, D. (2021b). Diagnostic feature training improves face matching accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition.* https://doi.org/10.1037/ xlm0000972

Tuddenham, W. J., & Calvert, W. P. (1961). Visual Search Patterns in Roentgen Diagnosis. *Radiology*, *76*(2), 255–256. https://doi.org/10.1148/76.2.255

Vincente, K. J., & Wang, J. H. (1998). An ecological theory of expertise effects in memory recall. *Psychological Review*, *105*. https://doi.org/10.1037/0033-295X.105.1.33

Watson, T. L., & Robbins, R. A. (2014). The nature of holistic processing in face and object recognition: *Current opinions. Frontiers in Psychology, 5.* Https://doi.org/10.3389/fpsyg.2014.00003. (n.d.).

CHAPTER 1: Relationships Between Expertise and Distinctiveness: Abnormal Medical Images Lead to Enhanced Memory Performance Only in Experts.

Abstract

Memories are encoded in a manner that depends on our knowledge and expectations ("Schemas"). Consistent with this, expertise tends to improve memory. Experts have elaborated schemas in their domains of expertise, allowing them to efficiently represent information in this domain (e.g., chess experts have enhanced memory for realistic chess layouts). On the other hand, in most situations people tend to remember abnormal or surprising items best – those that are also rare or out-of-the-ordinary occurrences (e.g., surprising – but not random – chess board configurations). This occurs, in part, because such images are distinctive relative to other images. In the current work, we ask how these factors interact in a particularly interesting case: the domain of radiology, where experts actively search for abnormalities. Abnormality in mammographs is typically focal but can be perceived in the global 'gist' of the image. We ask whether, relative to novices, expert radiologists show improved memory for mammograms. We also test for any additional advantage for abnormal mammograms that can be thought of as unexpected or rare stimuli in screening. We find that experts have enhanced memory for focally abnormal images relative to normal images. However, radiologists showed no memory benefit for images of the breast that were not focally abnormal, but were only abnormal in their gist. Our results speak to the role of schemas and abnormality in expertise; the necessity for spatially localized abnormalities vs. abnormalities in the gist in enhancing memory; and the nature of memory and decision making in radiologists.

Introduction

Our ability to remember information is deeply dependent on our existing knowledge structures, or schemas (Bartlett, 1932, Hintzman, 1986). Even superficially identical information is better remembered if it is integrated into a set of knowledge rather than simply seen as arbitrary. For example, people are better at remembering that someone *is* a baker than that someone's *name* is Baker, because the profession baker activates a rich set of meaningful associations that the name Baker does not (McWeeny, young, Hay, & Ellis, 1987); and people remember visual images better if they recognize them as face than if identical images are not recognized but seen as meaningless texture (e.g., Brady, Alvarez, & Störmer, 2019).

Different people have different knowledge and schemas, in part based on their expertise, and this has consequences for memory: Imagine after playing a round of chess, you are asked to recreate the board from some critical moment in the game. For most people, this task would prove very difficult. However, if you were a world-class chess player, this might be quite easy. Becoming an expert in a domain such as chess changes our memory for items in that domain of expertise (Chase & Simon, 1973; de Groot, 1946), allowing us to store more information as long as this information is consistent with the expectations we have formed as a result of our expertise (Gobet & Simon, 1996).

A large literature is devoted to quantifying memory benefits in experts compared to novices (e.g., Ericsson & Kintsch, 1995; Engle & Bukstel, 1978; Vincente & Want, 1998; Gobet & Sinon, 1996). For example, car experts can remember more car images in visual working memory (Curby, Glazek, & Gauthier, 2009); baseball experts can remember more baseball-related information in long-term memory (Voss, Vesonder & Spilich, 1980); and expert radiologists have better long-term memory for mammograms – but not natural scenes or real-world objects – compared to controls (Evans, Cohen, Tambouret, Horowitz, Kreindel, & Wolfe, 2011).

Why do experts show this increase in memory performance for their domain of expertise? In the literature on expertise, many authors posit that memory improvement occurs because existing knowledge allows experts to know what variation to expect for information in an expert's domain (e.g., Vincente & Want, 1988). That is, existing schemas make the relevant part of the information predictable and thus easier to encode and remember (Graesser & Nakamura, 1982). Thus, in many ways, memory benefits in experts may be considered a manifestation of a broader phenomenon where information that is understood as meaningful -- and thus integrated into a schema -- is easier to correctly recognize or recall (Bartlett, 1932). For experts, there may simply be a wider variety of meaningful concepts and schemas, resulting in a richer ability to understand and remember stimuli in their domain of expertise (e.g., Ericsson & Kintsch, 1995). This is sometimes know as an organizational processing account of expertise: that experts can have improved memory because they are better able to chunk this information and otherwise create effective knowledge structures (Ericsson & Kintsch, 1995; Rawson & Van Overschelde, 2008).

Is better organization the sole reason for better memory in experts? Beyond schemas and knowledge organization, experts in some domains – particularly those where the expertise is more perceptual, like radiologist looking at mammograms or car experts focusing on the details of cars – may have developed specialized processing mechanisms for their domain of expertise which take advantage of the way stimuli vary in that domain. For example, experts in some domains employ more holistic processing strategies for objects of their expertise (Watson & Robbins, 2014; Richler, Wong, & Gauthier, 2011; Gauthier, Skudlarski, Gore, & Anderson, 2000; Gauthier, Tarr, Anderson, Skudlarski, & Gore, 1999; Bilalic, Langner, Ulrich, & Grodd, 2011). Enhanced perceptual expertise may allow experts to process more information about an item even in the same amount of time, and lead to richer memory traces (Ericsson & Kintsch, 1995).

In addition to building richer knowledge structures and better perceptual encoding, there is a third factor that could explain experts' improved memory performance in domains of expertise, which has often been overlooked in studies of memory: increased distinctiveness of items when they are items of expertise (Rawson & Van Overschelde, 2008). In contrast to views that claim memorability is an intrinsic aspect of a stimulus (e.g., Bainbridge, Isola & Oliva, 2013), a significant literature argues instead that the critical driver of how memorable an item is in a given context is its distinctiveness from other items currently being stored in memory. Imagine, for example, you are given a list to remember that has 30 animal names and also the word 'bread' on it. People tend to remember this distinctive word ('bread') most accurately – and this is true even if it appears first on the list, so its unique status is not yet known and it is not differentially attended or processed (Calkins, 1984; Hunt, 2006). Memory models naturally predict this effect because most memory models propose that memory is strongly limited by interference at retrieval, and having more unique features allows easier retrieval (e.g., Shiffrin & Steyvers, 1997).

This is broadly consistent with the idea that abnormal or schema-inconsistent items tend to be *better* remembered than expected, schema-consistent items (Friedman, 1979, McDaniel & Einstein, 1986; Light, Kayra-Stuart, & Hollander, 1979; Pedzek, Whetstone, Reynolds, Askari, & Dougherty, 1989; Hollingworth & Henderson, 2003). For example, people tend to better remember unexpected aspects of images (Friedman, 1979).

How does such distinctiveness interact with expertise? For experts, many items may be unique from other items in a set in a way that would not be noticed by non-experts, thus enhancing memory for those items as they would then be more unique in the set for experts than non-experts (Rawson & Van Overschelde, 2008).

In summary, experts are often better at accurately recognizing or recalling information in their domain of expertise. This can arise from at least 3 factors, each of which has been independently studied: experts may have changed perceptual processing strategies; may benefit from general usage of schemas to organize memory; or may benefit from increased distinctiveness of items in memory. However, the way these effects interact has rarely been studied, and many have been studied primarily in domains with limited or no perceptual expertise available (e.g., in word lists).

The Current Experiments: Memory for Mammographs in Novices and Expert Radiologists

To understand how expertise effects memory, and how each of these three factors may play a role, the current experiments ask how expertise affects memory for mammograms (comparing novices and expert radiologists), and test whether expert radiologists have better memory for abnormal images (i.e., cancerous mammograms), when compared to normal images (i.e., noncancerous mammograms). While for normal mammograms, perceptual encoding benefits, schemas, and distinctiveness all likely plan a role in expert's memory, abnormal mammograms provide a unique case study. Abnormal mammograms do not violate a radiologists' schema (as they are trained to look for abnormalities), but abnormal cases do provide distinctive retrieval cues (e.g., this mammogram has calcifications in this location) which would not be available to non-experts who have no idea that those little white spots are significant. Nor would these cues be available in normal mammograms. Abnormal mammograms therefore present an interesting case; they are schema-consistent, while also providing a unique window into the role of distinctiveness in expert's memory.

To measure memory performance, we will use Receiver Operating Characteristic (ROC) analysis to take into account the possibility of differential false alarms and differential response criterion, which is critical to understand whether any effects we observe are truly changes in memory strength. We predict that experts will have improved performance compared to non-experts for both normal and abnormal mammograms because of their perceptual expertise and because they have developed schemas over time to represent these complex images. We also predict that abnormal items might show even more benefit for radiologists compared to non-experts because for radiologists and radiologists alone, these images have unique and distinctive retrieval cues available.

We focus on radiologists' memory for mammograms for two reasons: First, search for signs of breast cancer involves a usefully specific perceptual expertise. For instance, only 2-3 kinds of local abnormalities are typically present in abnormal mammograms, and radiologists have significant perceptual expertise whether looking at normal or abnormal medical images.

Second, there are two senses in which a mammogram might be considered "abnormal." 1) It could contain a focal abnormality. In our study, these are masses or architectural distortions that are subsequently proven to be malignant. 2) Given a mass (for example) in one breast, the other breast could be considered abnormal in the sense that the image comes from a patient with cancer. We assess the impact of each of these two kinds of abnormality on memory. Note that a mammogram might be considered ‘abnormal’ if it showed a benign mass. We did not use such stimuli in this study.

Radiologists are explicitly trained to recognize an image as abnormal if they detect the presence of a visible, localized abnormality like a mass or calcification. In addition, recent research has shown that, if asked in an experimental setting, radiologists have an ability to detect a “gist” of abnormality in the breast contralateral to the lesion. They perform at above chance levels when asked to categorize images as coming from normal or abnormal patients (Evans, Haygood, Cooper, Culpan, & Wolfe 2016). In other words, this study suggests that radiologists do not always need to see a localized physical lesion to know that an image is abnormal. This global signal of abnormality is relatively subtle. More importantly, for present purposes, work on this gist signal is new enough that most radiologists are unfamiliar with the concept. Thus, any impact on memorability could be considered to be the result of an implicit effect of abnormality.

Published studies of the gist of abnormality have involved giving radiologists only a brief (250-500ms) glance at the image. While this seems sufficient for expert radiologists to gain some evidence of abnormality, it remains unknown whether this ability impacts radiologists’ memory for normal versus abnormal images.

To summarize, the questions guiding this experiment are the following: Do radiologists show improved memory performance for abnormal images compared to normal images? If so, does global gist produce enhanced expert memory for images of the breast contralateral to the breast that contains focal signs of cancer? Alternatively, does any abnormality advantage in memory depend upon having a focal abnormality that can draw spatial attention?

Experiment 1 is a baseline study with novice observers, whose performance can be compared to radiologist performance in Experiment 2. In addition, Experiment 1 allows us to determine whether our stimulus set contains images that are memorable regardless of expertise. Experiment 2 assess memory performance in expert radiologists. To anticipate our results, Experiment 1 reveals patterns in our image set that we take into account in Experiment 2. In Experiment 2, we find a large memory benefit for radiologists relative to novices as well as an abnormality advantage in radiologists for focal abnormalities. We find no evidence that experts make use of a non-focal gist of abnormality either in judgment or memory.

Experiment 1: Novices

Experiment 1 was conducted using novice (non-radiologist) observers. The design, number of observers, exclusion and analysis plan for this experiment were preregistered (URL for this experiment: http://aspredicted.org/blind.pho?x=xr3843).

In this experiment, novice observers viewed a series of mammograms and judged whether each case was normal or abnormal and whether they remembered seeing the image earlier in the experiment. We would expect both judging whether an image is normal or abnormal as well as remembering the images to be difficult, as this task is designed for expert radiologists. However, novice performance provides a useful baseline for comparing radiologist performance, and provides a baseline of memory in novice observers. In particular, the results of this experiment can indicate if particular images are particularly distinctive in the absence of any mammographic expertise.

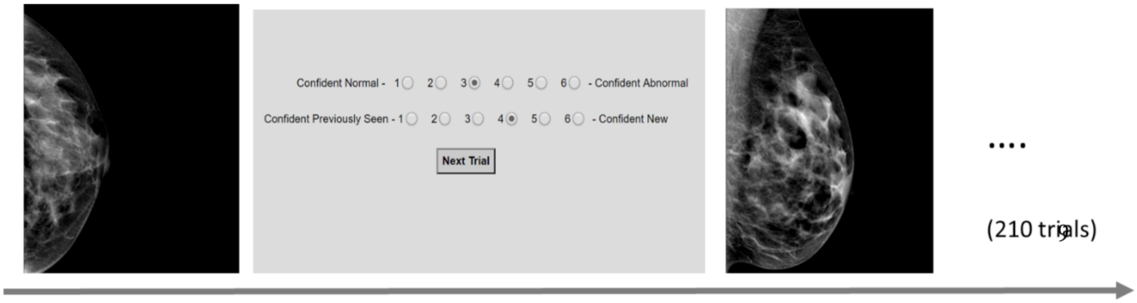
Method

Sixty participants (23 female participants, mean age 38 years) were recruited for this experiment through Amazon's Mechanical Turk, which offers monetary compensation for participation in online tasks. Mechanical Turk workers are reasonably representative of the American adult population (Berinsky, Huber, & Lenz, 2012; Buhrmester, Kwang, & Gosling, 2011; Difallah, Filatova, & Ipeirotis, 2018), and provide data that is comparable to data obtained when participants are tested in experimental psychological laboratories (e.g., see Brady & Alvarez, 2011 for a comparison in a visual memory context). All participants gave informed consent, were compensated at a rate of approximately $10/hour, were located in the United States, and had a Hit Approval Rate greater than 95%. Informed consent procedures were approved by the Institutional Review Board of the University of California, San Diego.

Participants viewed single breast mammograms in this study. The stimulus set consisted of 80 abnormal (cancerous) cases and 40 normal (non-cancerous) cases. All images were deidentified. All images were pre-classified by a group of trained radiologists who did not participate in the study. Normal images were noncancerous and did not contain benign lesions. Abnormal images consisted either of histologically verified malignant masses or architectural distortions (see Evans et al., 2016, for a more detailed description of this stimuli set). Half of the abnormal images contained a visible abnormality (i.e. a lesion was present) and half were images of the breast contralateral to the breast with the lesion (i.e. still an abnormal case, but with no focal indication of that abnormality). Thus, the entire set consisted of 40 normal images, 40 focal-abnormality images (herein referred to as abnormal) and 40 non-focal abnormality images (images contralateral to the breast with the focal abnormality), herein and henceforth referred to as contralateral-abnormal. Each image subtended approximately 16 x 20 degrees of visual angle at an estimated viewing distance of approximately 60 cm from the screen.

On each trial, one image was present for 3 seconds, followed by a new screen containing response questions. The mammogram was randomly chosen to be either normal, abnormal, or contralateral-abnormal. Critically, each image was also either a new image (presented for the first time in the experiment), or a repeated image from 3 trials back or 30 trials back (3-back and 30-back, respectively). Of the images that were later repeated, 50% were repeated at 3-back, and 50% were repeated at 30-back. In total, with repetitions, there were 210 trials: 120 new images (40 per condition), plus 90 repeat images (30 per condition, split evenly between 3-back and 30-back).

After being displayed for three seconds, each image was immediately followed by two response questions: (1) Was the image abnormal or normal? (2) Have you seen this image before? Using a standard computer mouse, participants were told to indicate their level of confidence on a six-point rating scale rating from confident yes/abnormal to confident no/normal (Figure 1.1). We collected confidence ratings instead of simple yes/no answers to allow for ROC analysis. There was no time constrained imposed on responding. The initiation of the next trial was contingent on answering both questions of the current trial.


Confident Normal - 1○ 2○ 3● 4○ 5○ 6○ - Confident Abnormal

Confident Previously Seen - 1○ 2○ 3○ 4● 5○ 6○ - Confident New

Figure 1.1. *Method.* N=60 non-expert novice participants rated a sequence of 210 images on normal/abnormal and old/new. Images could repeat either after 3 or 30 subsequent images and be either normal, abnormal, or contralateral-abnormal.

Before the experiment began, participants were presented with instructions and several demographic questions (Gender; Age; 'Are you a radiologist?'; 'Do you have a job where you read medical images, (i.e. tech, medical physicist)?'). Instructions were written for a novice population with no medical training. For novice participants, abnormal cases were broadly defined as "images that might contain lesions, or cancer, or otherwise might be something worthy of follow-up if you were a radiologist."

Our exclusion criteria and analyses were decided in advance (see preregistration, above). Individual trials were excluded if participants took less than 1500ms or more than 15,000ms to respond (based on pilot data). Participants were excluded if they took less than 15 minutes (0 excluded) or more than 1 hour to complete the study (3 excluded). Radiologists were excluded (1 excluded) as were those with other prior experience reading medical images (0 excluded). Participants were also excluded if they had more than 80% identical responses (e.g., picked the exact same answer on nearly every trial; 1 excluded) or had more than 30% of trials excluded on the basis of the reaction time criteria (1 excluded).

After applying these a priori exclusion criteria, 7 participants were excluded from analysis, leaving a final sample of 53 participants.

We first analyzed the confidence ratings of classifying an image as abnormal or normal. We subsequently analyzed the confidence ratings representing memory for images. In order to do this, we conducted ROC analysis for the 3-back and 30-back as a function of image type (normal/abnormal/contralateral-abnormal). We also generated ROCs for the normal/abnormal judgments. ROCs were summarized by area under the curve (AUC) and compared using t-tests. As noted, we are interested in whether, within the group of novice participants, there is a benefit in memory performance for any type of image (e.g., as judged by normal vs. abnormal AUC). Since the novices lack medical experience, any such effect would give us insight into the nature of the image set (i.e., memorability or distinctiveness). Finally, we conducted image similarity analyses to quantify how image differences might be influencing these results.

Image Similarity Comparison: Because normal, focally abnormal and contralateral-abnormal images are necessarily different image sets, it is useful to compare how distinctive each set of images is from all the other images in order to look at the effect this has on memory. One way to accomplish this is to have individuals give similarity ratings between images. However, this would require 120*120 = 14,440 similarity ratings. Instead, to streamline the process, we relied on previously established computer vision techniques designed to give similarity measurements for natural scenes. Specifically, we conducted a gabor wavelet pyramid (GWP) analysis, which computes features of the images and compares them (Kay, Naselaris, Prenger, & Gallant, 2008; Greene, Baldassano, Esteva, Beck, & Fei-Fei, 2016). To assess the level of similarity in the different image types, the

GWP represents each image as the output of a bank of multi-scale Gabor filters. Prior work has shown that these features can successfully model object representation in early visual areas (Kay et al., 2008). Following the exact procedure and parameters provided by Greene et al. (2016), each image was converted to grayscale, down sampled to 128 by 128 pixels, and represented with a bank of Gabor filters at three spatial scales (2, 6, and 11 cycles per image with a luminance-only wavelet that covers the entire image), four orientations (0, 45, 90 and 135 degrees) and two phases (0 and 90 degrees). This gave a set of features for each image, which we then compared to all 120 images to compute a distance / dissimilarity score by computing the dot product of each images features to each other images after subtracting the mean across images and normalizing the feature vectors to unit length.

Results (Experiment 1: Novices)

Performance on the classification task: First, we looked at how confident novices were at classifying an image as either normal or abnormal (Figure 1.2). We found a significant difference between normal and abnormal images ($t(52) = 4.78$, $p<0.001$), but not between normal and contralateral-abnormal images ($t(52) = 1.94$, $p>0.05$).

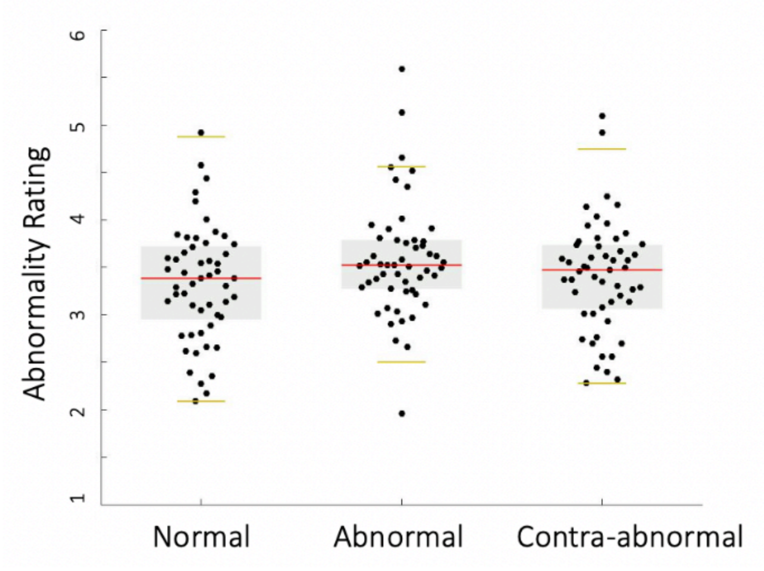


Figure 1.2. Classification Task. Overall performance of novices on labeling an images as normal or abnormal. The confidence rating scale is now plotted on the y-axis. Each point in the plot represents the rating for a particular image. We found a significant difference in confidence in classifying normal vs. abnormal images, which seems to be driven by a few salient abnormal images. Novices are not confident in distinguishing between any image type (most responses tend to be in the middle of the confidence scale no matter the image type). Error bars represent standard error of the mean.

While participants did not have training to distinguish normal from abnormal medical images, a small number of images in the set are extremely saliently abnormal (i.e., a single bright white spot would look questionable even to novice viewers). Looking at ratings by image (Figure 1.3) reveals that these images are largely responsible for the significant difference between normal and abnormal ratings. In short, for at least a small subset of images, even novice participants can notice the abnormality, leading to above chance classification performance broadly. But for most images, novices seem to have little information about normality vs. abnormality.

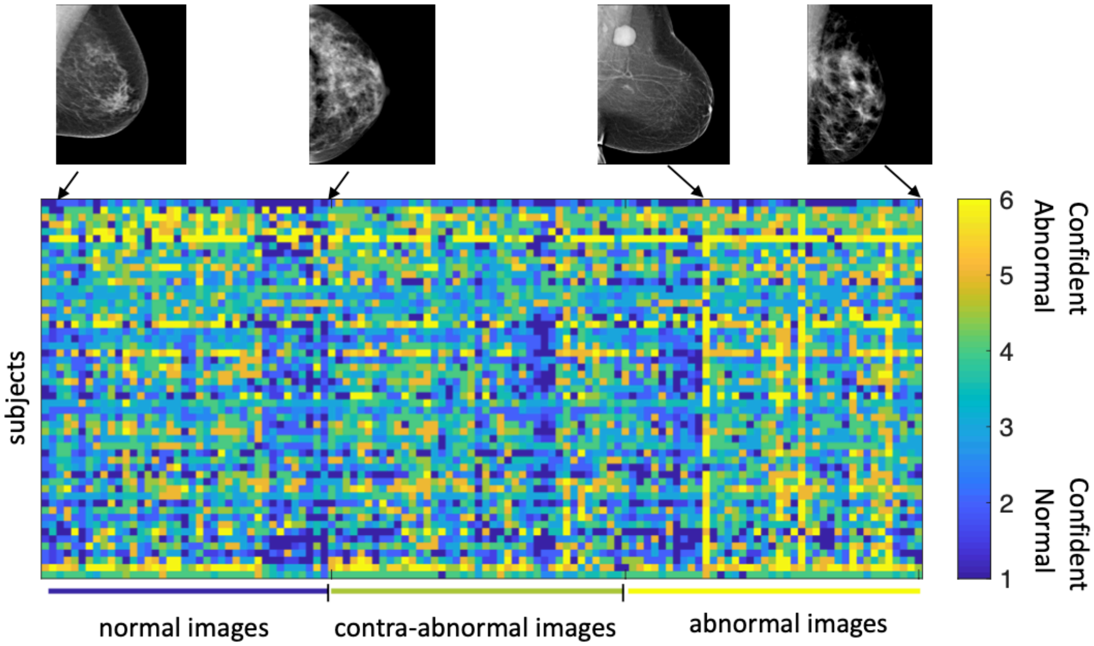


Figure 1.3. Image ratings in the classification task. Example images and their confidence ratings for each participant in the classification task. As can be seen with the third pictured image, most participants rated this as abnormal with high confidence. Altogether, the two or three brightly striped vertical lines in the abnormal images set indicate that those and only those images were reliably rated as abnormal by a large majority of participants.

Note that the y-axis in Figure 1.2 represents the confidence ratings for novices. It is clear that the novices are generally not confident in distinguishing any image type, with average responses tightly clustered near the middle of the rating scale for all conditions. Another way of visualizing this data is on an ROC curve (Figure 1.4), where novices fall almost on top of the dotted diagonal line indicative of chance performance, with an AUC of only 0.54 (where 0.50 is change and 1.0 is perfect). Although, as stated above, this difference from chance is highly reliable across participants ($t(52) = 4.21$, $p<0.001$), largely because of the few images that participants could all reliably classify.

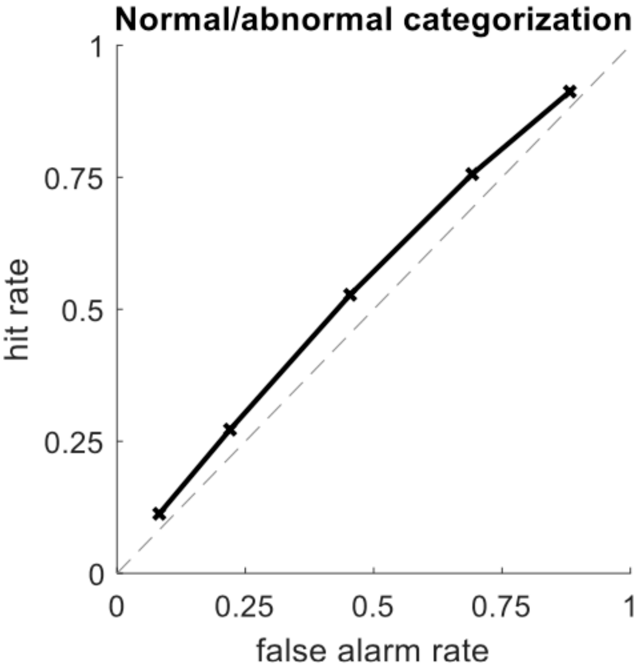


Figure 1.4. ROC for normal/abnormal categorization. Novices are very close to the diagonal line representative of chance performance, indicating that they do not perceive a strong difference between normal and abnormal images. The significant effect is driven by a select few salient images (see Figure 1.3).

Memory for abnormal images: Figure 1.5 shows the ROCs for the 3-back and 30-back memory tasks. Since novices were not, for the most part, able to perceive contralateral-abnormal images as different from normal images in the classification task, we focused exclusively on memory differences between normal and abnormal images. Overall, independent of image type, and as expected, novices have better 3-back memory (averaged AUC of 0.70 for detecting 3-backs) than 30-back memory (averaged AUC of 0.64 for detecting 30-backs), t(52)=6.59, p<0.001. Interestingly, breaking down performance across image conditions reveals that novices show a small normality benefit: they remember normal images better than abnormal images in both the 3-back condition and the 30-back condition, with only the 3-back yielding a significant difference. We found an AUC benefit of 0.069 for normal images at 3-back (t(52)=5.48, p<0.001) compared to abnormal, and an AUC difference of 0.026 for normal images at 30-back (t(52)=1.70, p=0.096) compared to abnormal.

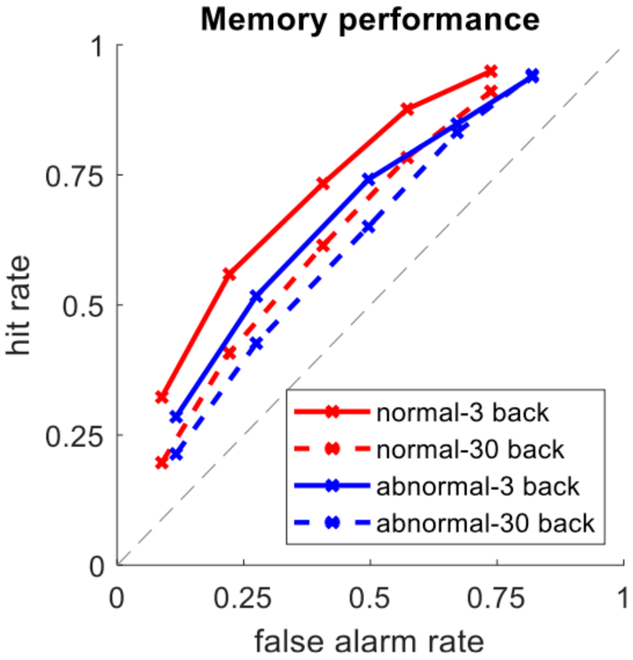


Figure 1.5. Novice performance on the memory task. As noted, the gray dashed line indicates chance and more bowed out curves represent better memory performance. Novices had stronger memory for images in the 3-back condition than in the 30-back condition. Novices also show a small effect of normality, with memory for normal images being better than for abnormal images in both 3-back and 30-back conditions.

Given the weak performance at discriminating normal from abnormal images, it is rather surprising that normality had any effect. Therefore, we examined the data for evidence of more basic effects of visual similarity. We found that the lower memory performance in the abnormal conditions was largely driven by an increased false alarm rate in the contralateral-abnormal and abnormal image sets. Here we are classifying as "new" all images with a confidence rating > 3. This is consistent with an image similarity account in which novices would be more likely to false alarm to new images in the contralateral-abnormal and abnormal conditions simply because these images are more similar to one another than images in the normal set (as predicted by summed similarity accounts of memory, e.g., Nosofsky, 1991). In other words, if the normal images were somewhat more dissimilar to each other compared to the other images, this could explain why novices have somewhat better memory for the normal condition (i.e., it is easier to determine if an image of a dog is new if that dog is presented in a series of different animals than if it is presented in a set of similar dogs. Obviously, the similarity effects in our stimuli are much smaller). We test this hypothesis next.

Similarity Matrix – Gabor Wavelet Pyramid Analysis: We tested this image similarity hypothesis by measuring similarity between our images as described in the Methods (Kay et al., 2008; Greene et al., 2016). We found increased *dissimilarity* among normal images relative to contralateral-abnormal and abnormal images (Normal = 0.174; Abnormal = 0.139; Contralateral-abnormal = 0.133). In other words, normal images were more different, on average, from one another (and thus more discriminable in memory) than either abnormal or contralateral-abnormal images. This is consistent with the hypothesis that the small differences in memory favoring normal images is driven by image similarity differences across sets. Thus, the small normality benefit found in the current study is likely a result of image similarity. Critically, this can provide a useful baseline for considering memory for the same images in expert radiologists in Experiment 2.

Experiment 2: Radiologists

Experiment 2 was the same as Experiment 1, except conducted on radiologist observers.

Method

Thirty-two expert radiologists (14 female participants, average age = 49 years) were recruited during the 2018 Radiological Society of North America (RSNA) conference in Chicago, Illinois. All radiologists gave informed consent and were not compensated beyond being entered into a lottery for a $500 gift card. Informed consent procedures were approved by the Institutional Review Board of the University of California, San Diego.

Data from participants would have been excluded if they took less than 15 minutes or more than 1 hour to complete the study, had more than 80% identical responses or had more than 20% of trials excluded. Under these guidelines, no radiologists were excluded from analysis, leaving a final sample of 32 participants.

The stimuli and experimental design were the same as described in Experiment 1. The main procedural difference was that the experiment was conducted at the RSNA conference where the experimenter explained the instructions in person. Unlike in Experiment 1, in Experiment 2, we gave more general instructions, asking for any abnormality rather than specifically asking participants to look for focal lesions or cancer: "For each image, please judge whether the image is abnormal or normal, and whether you have previously seen it during the course of the experiment."

Results

In this section, we compare the performance of expert radiologists to the performance of novice participants in Experiment 1. In particular, we investigate how non-experts compare to experts' judgments of image classification (i.e. normal vs. abnormal), and critically, whether experts show differential memory for abnormal versus normal images. While analyzing expert performance, we take into account idiosyncrasies in our image set that we learned from Experiment 1, such as that our normal images are more dissimilar and therefore inherently slightly more memorable.

Performance on the classification task: Similar to Experiment 1, we first analyzed performance on the classification task by looking at the confidence ratings of classifying each image as either normal or abnormal. How good are radiologists at simply distinguishing abnormal from normal images? Unsurprisingly, radiologists are very good at distinguishing abnormality (Figure 1.6A). Radiologists were significantly more confident that an abnormal images was abnormal instead of normal ($t(31) = 13.17$, $p < 0.001$). Figure 1.6B shows the ROC curve for distinguishing focal-abnormal images from normal images in radiologists. ROCs were summarized by area under the curve (radiologist AUC = 0.72; recall that novice AUC = 0.54). As noted in Experiment 1, controls are close to the diagonal line indicative of chance, whereas radiologists elicit a typical curvilinear ROC indicative of a perceived (and significant) difference between normal and abnormal images with an AUC well above chance ($t(31) = 19.8$, $p<0.001$).

Next we looked at if radiologists could detect abnormality in the contralateral-abnormal images. There was not a significant difference between the normal and contralateral-abnormal image conditions ($t(31) = 0.43$, $p = 0.67$). In the original study of Evans et al. (2016), they found an effect of abnormality in the gist information (i.e., in a very short presentation time of ~250ms). Our instructions and stimulus set may have biased participants against reporting contralateral images as abnormal. In a set of images that include visible lesions (the abnormal cases) and in the absence of an instruction to look for asymptomatic images from symptomatic patients (the contralateral cases), it is, perhaps, not surprising that radiologists reserved their abnormal ratings for the abnormal cases with lesions. Furthermore, it is possible that our instructions could have primed radiologists to look for both benign and malignant lesions, although no benign lesions were present in the current study. Future studies could investigate the effects of instruction on this task. Recall, however, that our interest in the present experiment is in radiologists' memory for these images. Contralateral- abnormal images, for instance, might still be remembered better if their vaguely suspicious appearance caused radiologists to devote more attention to them.

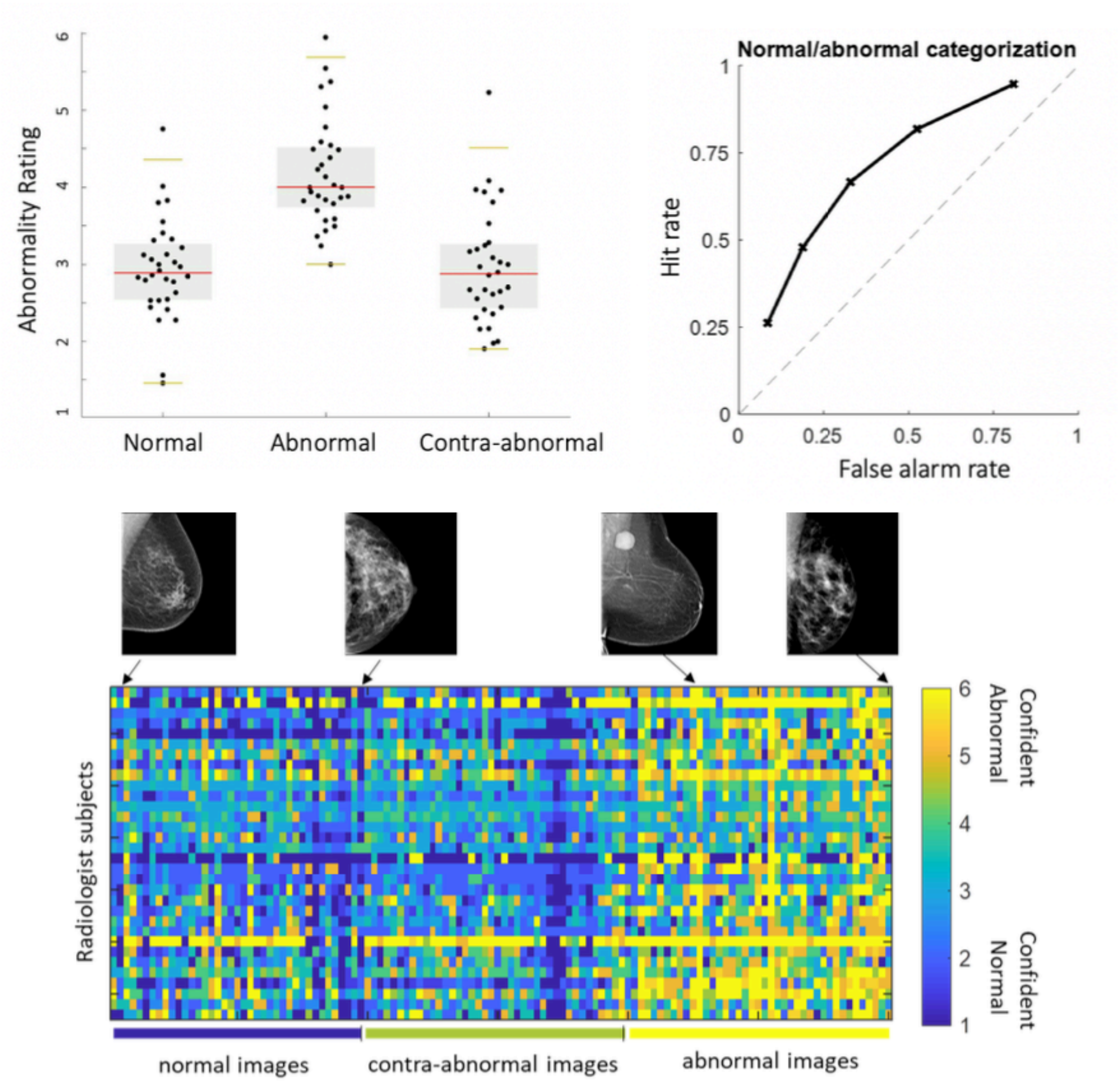


Figure 1.6A (top left): Classification task: Overall performance of radiologists on labeling an image as normal or abnormal. Once again, each point in the plot represents the average rating for a particular image. Radiologists clearly distinguished abnormal from normal images, but they did not distinguish between contralateral-abnormal and normal images. **1.6B (top right):** ROC depiction of performance for labeling an abnormal image as abnormal instead of normal (ignoring contralateral-abnormal images). **1.6B (bottom):** Classification by image. Unlike novices, experts reliably classify most of the abnormal images as abnormal and most of the normal images as normal, with performance not largely driven by any particular subset of images.

Figure 1.7 shows radiologist performance on the memory task. Radiologists have better memory for abnormal images in both memory conditions, but the advantage for abnormal images is only significant in the 30-back condition (t(31) = 2.86, p = 0.008, AUC difference – 0.051). We found an AUC advantage of 0.02 for abnormal images at 3-back. Although this was not significant (t(31) = 1.62, p = 0.12), it follows the same trend as the 30-back condition.

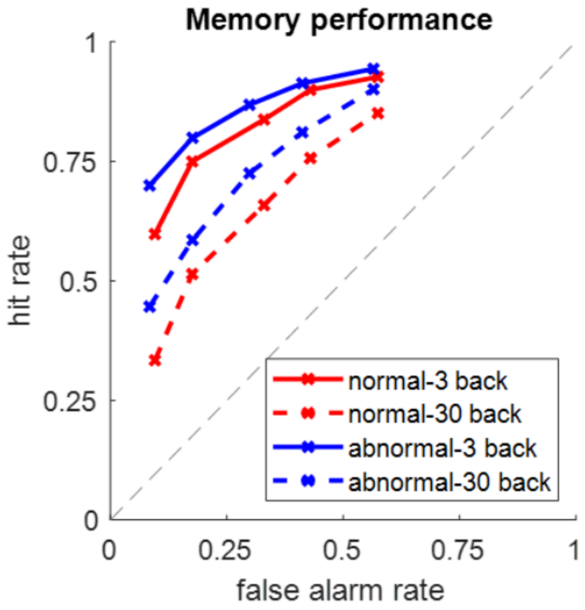


Figure 1.7. Radiologist performance on the memory task. Radiologists have better memory for abnormal images in both of the memory conditions. However, only memory at long delays (30-back) was significant.

Radiologists showed no memory benefit for the contralateral-abnormal images, even at long delays (p=0.24). Since radiologists were not able to distinguish between contralateral-abnormal images and normal images in the classification task, this result might be expected; though, recall that we were looking for evidence that an implicitly recognized abnormal gist might enhance memory. That is not what we found. Overall, independent of image type, radiologists have better memory at 3-back (averaged AUC of .852 for detecting 3-backs) than 30-back (averaged AUC of .752 for detecting 30-backs) for medical images. Why are radiologists better at 3-back than at 30-back? While it seems clear that this difference largely reflects typical effects of forgetting and interference (e.g., Wixted, 2005), it is also possible that observers would be more likely to 'catch on' to the presence of 3-back rather than the 30-back repetitions. If so, they might adopt a strategy that prioritized the 3-back task. However, given that the 3-back and 30-back tests were equally likely and equally distributed throughout the task, and that observers consistently said they remembered seeing mammograms from 30 images back (and therefore were distinctly aware that 3-back wasn't the only n-back test present), it seems unlikely that observers would transition to a strategy that only prioritized 3-back memory task. Taken together, these results suggest that experts have better memory overall at 3-back than at 30-back, but that a memory benefit for abnormal images compared to normal images is significant only at 30-back.

In recognition memory studies, it is almost always found that ROCs are not consistent with an equal variance signal detection model (e.g., Egan, 1958; Wixted, 2007). One way to look at this is to convert the hit and false alarm rates to z-scores and to plot zROC functions. On a zROC graph, equal variance produces data with a zROC slope of 1.0. Instead, as is typical in recognition memory tasks, the slopes of our zROCs were reliably below 1.0 in 3 of the 4 memory conditions. We fit a linear mixed effect model with slope and intercept as random per subject factors (mean slope[M]=0.68 for 3-back for normal images, difference from 1.0: $p<0.001$; M=1.05 for 30 back for normal images, not different from 1, $p=0.60$; M=0.52 for 3 back for abnormal images, $p<0.001$; 0.82 for 30 back for abnormal images, $p=0.005$). Collapsing across all conditions, thus allowing the slope to be more reliably estimated, the mean zROC slope was 0.68, significantly different from 1.0 ($p<0.00001$). Taken together, then, the ROCs we observed in memory were inconsistent with an equal variance signal detection model and consistent with an unequal variance model, potentially due to variation in memory strength between different items. This is typical of recognition memory and the reason that collecting confidence judgments and performing ROC analysis is necessary in order to assess memory strength. Simple d', in this context, does not properly account for response criteria differences (e.g., Dougal & Rotello 2007).

Recall from the similarity analysis in Experiment 1 that the normal images in our data set are less similar to each other than the abnormal images, and thus memory for normal images should be better than abnormal (as it was in novices). In fact, it is memory for the abnormal images that is better in radiologist observers. This suggests that the effect of expertise more than compensates for differences between the stimulus categories in image similarity. To see what the effect of abnormality is, independent of baseline image similarity differences, we can compare radiologists' memory performance to novices' performance with the same images. To do this, we compare the benefit – in terms of AUC of the ROC – for radiologists relative to controls in each condition. Doing so reveals a significant abnormality benefit at both 3-back ($t(31) = 6.67$, $p < 0.001$) and 30-back in expert radiologists ($t(31) = 4.43$, $p < 0.001$), where, taking performance after baselining relative to the performance of novice participants, radiologists were specifically better at remembering abnormal images (Figure 1.8).

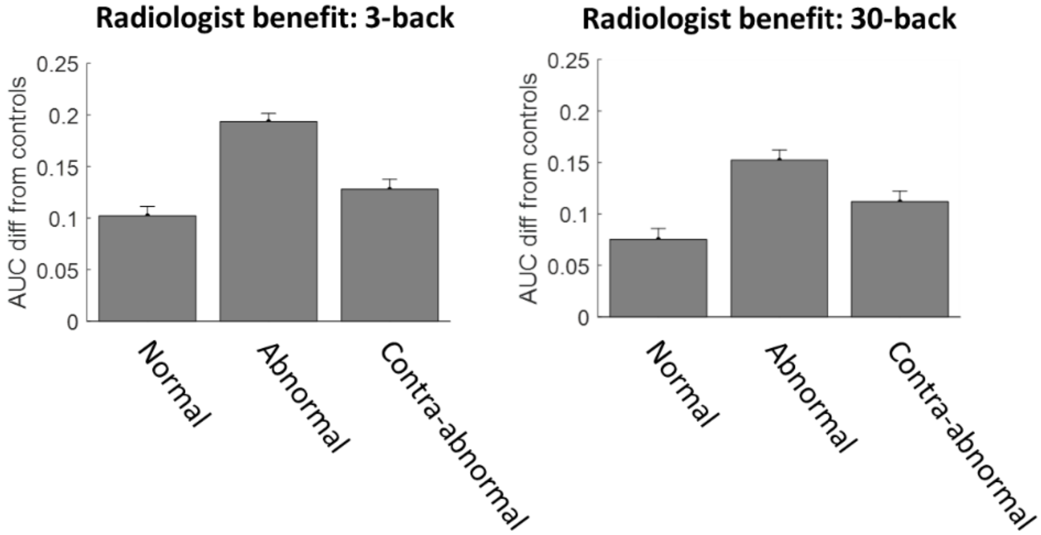


Figure 1.8. Using novices as a baseline to account for image similarity, there were robust abnormality memory benefits for radiologists at both 3-back and 30-back.

Due to the structure of this experiment, designed to probe memory, each item in the memory set has two classification ratings (for normal/abnormal). Thus, while we set out to probe memory, the experiment also makes it possible for us to combine both ratings in order to examine whether there is a "crowd-within" effect in this situation (Vul & Pashler, 2008). The authors proposed the crowd-within as a variant for the "wisdom of the crowd." They found that averaging a single individual's response to repetitions of the same question led to better performance than single responses alone. This is what one would expect if a single judgment did not incorporate all of the information people could possibly have about a question. If this is true for assessments of mammograms by expert radiologists, we would expect that averaging a radiologist's ratings of abnormality from two exposures to the same mammogram should result in better accuracy than looking at either rating alone. Note that in this situation, however, unlike Vul and Pashler (2008), participants actually have additional information the second time: they get to see the image again before the second judgment. They are not just asked again. Thus, in this case, the crowd-within effect here could arise from actual new information being incorporated (e.g., the observer might scrutinize different parts of the image), rather than internal sampling.

We find a modest, but significant advantage to incorporating both judgments: averaging radiologists' responses from the first and second time that they saw an image resulted in slightly higher performance in the 30-back condition (AUC=0.745) compared to single item performance (AUC=0.716; t(31)=3.46, p=0.002) (Figure 1.9, left). The effect was not significant in the 3-back condition (Joint AUC = 0.712, single AUC = 0.705, t(31)=1.15, p=0.259). Unsurprisingly, this effect was not present in novices, since their performance was very poor on both responses (Figure 1.9, right; all p>0.10).

Thus, expert performance can be improved (albeit, rather modestly) by averaging more than one response. It remains to be seen whether this benefit would occur if radiologists were offered unlimited time to process each image, rather than the 3 seconds in the current study. The limited viewing time here may have particularly enhanced radiologists' ability to extract new information in the second viewing of the mammogram.

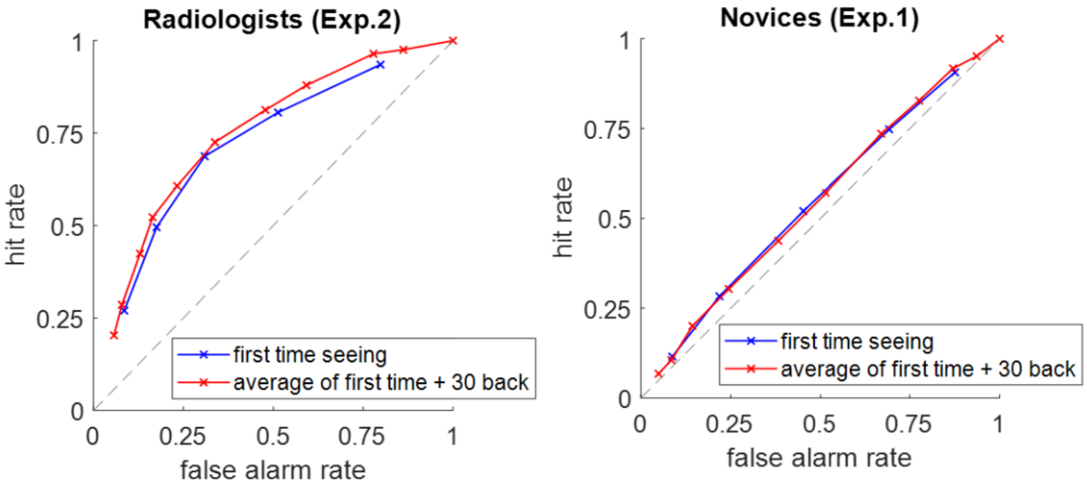


Figure 1.9. Crowd-Within Analysis: Left, Radiologists (Experiment 2): The blue line is the ROC for distinguishing focally abnormal mammograms vs. normal mammograms when the radiologists first see the image. The red line is the average of the first time seeing it and their responses seeing it at 30-back. Right, Novices (Experiment 1): Once again, the blue line is the ROC for distinguishing focally abnormal mammograms vs. normal mammograms when novices first see the image. The red line is the average of the first time seeing it and their responses seeing it at 30-back.

General Discussion

In the current study, we examined memory performance by non-expert novices and expert radiologists for normal versus abnormal mammography images as a case study in understanding the role of schemas, distinctiveness, and expertise in memory. To do so, we relied on ROC analysis, designed to properly measure memory independent of differences in response criteria and to take into account both enhanced memory for seen items as well as the possibility of false alarms.

First, we looked at how confident and how competent novice and expert observers were at classifying medical images as either normal or abnormal. Unsurprisingly, radiologists were much better than novices at this task. Novices did show some ability to distinguish abnormality, although this appeared to be largely the result of a few salient images.

Second, we examined our main question of interest: memory for the images. In Experiment 1, we examined memory for mammograms n novices, who have none of the expertise or schemas needed to processes these images. We found poor performance overall, as well as a small normality. Benefit in novice participants' memory, which could be explained by the greater image dissimilarity of normal images. Thus, Experiment 1 (on novices) gave us not only a baseline for memory performance, but also an understanding of the intricacies of our image set, showing that some abnormal images were quite salient, and that our normal images were more dissimilar from each other.

Even though the normal images in our set were more visually distinctive, in Experiment 2 we found that radiologists had better memory for abnormal images, and had far superior memory performance to novices. This gives insight into how expertise changes memory: not only enhancing the encoding of normal items, but also enhancing the distinctiveness of abnormal items. Thus, while experts might have access to perceptual encoding benefits, distinctiveness and/or schemas/chunking to enable them to outperform novices, our finding of an extra benefit of expertise for abnormal images is most consistent with a special role of distinctiveness. For experts, the abnormal images have unique features that make them distinct from other items in memory; whereas for novices, these features are not appreciated and so these images are just like any other image. For example, one possibility is that rather than encoding the entire image, in the case of abnormal images radiologists specifically encode the abnormality and not the rest of the image into memory.

This might reduce the load on memory for that image and might make the memory trace for that image more distinctive.

Broadly speaking, then, we find strong evidence for a role of schemas and distinctiveness in memory, even after taking into account false memory and the possibility of response criterion shifts: We find experts significantly outperform novices and that memory for abnormal cases with a visible, focal lesion is better than memory for other images. There was no evidence for a memory benefit for "abnormal" contralateral cases.

Measuring Memory: False Alarms and ROC Analysis

In the current studies, we used ROC analysis to examine memory. This is because, in previous work, it has often been unclear if benefits for schema-consistent information like those reported in experts are, in fact, improvements in memory as opposed to changes in response criteria. To determine if memory has actually improved, it is not adequate to simply find a reliable increase in the rate with which observers correctly report having been exposed to some piece of information (the true positive or "hit" rate). The observer could simply be saying "yes, I have seen it" more often. This would produce an increase in false positive (or false alarm) errors. In the context of memory research these false positive errors can be seen as a form of false memory. In theory, signal detection models and measures like d' can distinguish between these two, but in practice, the prerequisites for d' to properly adjust for response bias (equal variance; zROC slopes=1.0) are almost never present in recognition memory contexts, and were not present here. Thus, ROC analysis is needed to distinguish between the difference in the ability to remember as opposed to criterion shifts, which would reflect an increased tendency of observers to say that they remember (e.g., Wixted & Mickes, 2015).

Is false memory a true concern? In fact, previous work has found that organizing information in memory via schemas can have both positive and negative consequences – and in particular, does often increase false alarms, making it difficult to tell whether memory is genuinely improved. In particular, while greater understanding – as in expertise – may allow encoding of only the relevant details, reducing memory load, it may also cause us to falsely remember information that was not present (e.g., Owens, Bower & Black, 1979). For example, in recognition tests, people are more likely to false alarm to schema-consistent relative to schema-inconsistent lures. They would be more likely to falsely report seeing books in a graduate student's office than inconsistent objects like a piece of tree bark or a pair of pliers (Brewer & Treyens, 1981; Lampinen, Copeland, & Neuschatz, 2001). And while participants are more likely to correctly remember schema-consistent information in a briefly presented scene (Biederman, Mezzanotte, & Rabinowitz, 1982; Brewer & Treyens, 1981), they are also more likely to falsely remember such information (e.g., Pedzek et al. 1989; Hollingworth & Henderson, 2003).

Thus, measuring fully ROCs – rather than attempting to infer how response bias would change performance using measures like A', d' or hits minus false alarms – often reveals surprising answers about memory, particularly in situations like expertise and consistent/inconsistent items where it is known that both hit and false alarm rates are affected. For example, Dougal and Rotello (2007) used ROC analysis to show that the well-known effect of "improved memory" for emotional words compared to neutral words is a response bias effect, not a true difference in memory between the words. Similarly, Mickes, Flowe and Wixted (2012) showed in the domain of eyewitness memory that sequential line-ups, which reduce both false alarms and hit rates relative to simultaneous line- ups, are actually inferior to simultaneous line-ups, contrary to a large literature suggesting the opposite (e.g. Wells, Steblay, & Dysart, 2011), as the major ‘benefit’ arises simply from a response criterion shift, not a change in memory strength.

Thus, the current experiments provide unique evidence that expertise and distinctiveness that is apparent only to experts do, in fact, enhance memory – and that this is not just a response criterion shift.

What Explains Radiologists Outperforming Novices

Consistent with a wide variety of work on expertise, we find that expert radiologists outperform novices in remembering mammograms. One likely possibility is that this occurs because of experts knowledge about these images: they have relevant knowledge that allows them to understand these images in a way novices do not, and likely have perceptual expertise built into their visual system from years of experience (e.g., in the form of greater holistic processing; e.g. Richler, Wong, & Gauthier, 2011). In particular, for an expert, the abnormal images would have an added attribute (that mass, that calcification), learned over years of experience, that would help to distinguish the item in memory.

However, in the current study, we did not attempt to directly match our experts to our novices. Our novice pool was sampled from the internet, which is much more broadly representative of the demographics of the United States than an undergraduate population (e.g., Difallah, Filatova, & Ipeirotis, 2018), but still likely differs in a number of ways from our radiologists (in demographic and socioeconomic factors, as well as motivation to focus on mammogram images). Thus, Experiment 1 should be taken as only an approximate baseline: it revealed important image features in our stimulus set, and points to the possibility of strong expertise effects, but does not directly confirm these are based solely on knowledge rather than other factors.

Memory and Abnormality Judgments in Radiologists

Previous work has found mixed results when investigating memory improvements in radiologists. For example, Hardesty, Gannott, Hakim, Cohen, Clearfield, & Gur (2005) investigated radiologists' long-term memory for medical images presented months later and found that none of the radiologists remembered cases that they had read previously. Evans et al. (2016) found mixed results when investigating whether abnormality improves memory in expert observers, including radiologists. Our results provide context to these ambiguities, as they suggest that expert radiologists do have stronger memory for abnormal images even in a long-term memory setting and even when response bias is properly taken into account using ROC analysis. However, our long delays were only on the order of minutes, not months, and so it remains unclear how such advantages would last over long durations.

It is worth noting that in the classification task, radiologists performed on average much more poorly than would be expected of radiologists in the clinic with unlimited viewing time (d' = 2.5-3.0, as in D'Orsi et al., 2013). One reason for this might be that each image in our study was only presented for 3 seconds each. For instance, Evans, Georgian-Smith, Tambouret, Birdwell, & Wolfe (2013) showed radiologists only a brief glimpse of mammograms and varied timing from 250ms to 2000ms. The respective AUC's for radiologists in their experiment for 500ms, 1000ms, and 2000ms viewing times was 0.65, 0.66, and 0.72. In our experiment with a presentation time of 3000ms, we found an AUC of 0.72. Thus, our 3000ms presentations resulted in a similar level of performance to the 2000ms presentations of Evans et al. (2013), which, while well below what is expected with unlimited viewing time, is consistent with other studies and consistent with viewing time being the main constraint that lead to lower performance.

The “Crowd-Within” Effect in Radiologists

Because our study had radiologists answer the same classification question about an image multiple times, we looked at whether averaging radiologists’ responses when they judge the same image twice resulted in better performance (a “crowd-within” effect, Vul & Pashler, 2008). We found that radiologist performance improved when averaged across the same image twice compared to either response alone, but only in the 30-back condition and only modestly even then. This indicates that by the time radiologists were presented with the same image 30 images later, they gave a response that is somewhat independent of their first response. This suggests that, under the current experimental conditions, there might be information the radiologists are not using the first time they see an image – and that the opportunity to see the image again allows the radiologist to glean additional useful information. Future studies might determine whether such benefits persist when experts are given unlimited time to process the images as well as whether this effect can be made larger with an even longer delay between the first and second presentation of an image (as found by Vul & Pashler, 2008).

The “Gist” of Abnormality

Given the Evans et al. (2016) finding that there is a “gist of abnormality” present in the contralateral breast when no localizable abnormality is present, we were interested to know if these contralateral- abnormal images had any advantage over normal images in expert memory. We found no such evidence. In our experiment, we also found no difference in the classification of abnormality between contralateral-abnormal images compared to normal. While at first this might seem to contradict earlier work, there are a number of methodological differences that make it difficult to compare our results directly with Evans et al. (2016). It is possible that we did not find this result because we presented images for longer encoding time (3000ms). Typical stimulus exposure in mammogram "gist" studies has been less than a second; 500ms is typical. It is possible that presenting images for longer encoding times might actually obscure the gist information - overwriting an initial "gist" impression with more semantic or meaningful information. Recall, also, that our radiologists were not informed about gist and likely reserved their 'abnormal' ratings for cases where they could localize a lesion. It is possible that we would observe a contralateral- abnormal effect even at long encoding times if we explicitly directed participants to look for a more general abnormal texture or gist. Given these methodological differences, the current study cannot be readily compared to Evans et al. (2016). However, this seems to be a promising avenue for future work.

Conclusion

Using radiologists as a case study, we find an advantage for memory in experts as well as an advantage for abnormal images – even when properly measuring memory via ROC analysis. This is broadly consistent with the literature on schemas. Our findings have important implications for both applied fields that utilize expert intelligence in making inferential decisions as well as theoretical fields interested in how memory changes with expertise. In particular, understanding the structure of memory in experts is critical in situations where decisions need to be made by people who have significant expertise.

Chapter 1, in full, is a reprint of the material as it appears in Memory and Cognition, 2021, co-authored with Dr. Jeremy M. Wolfe and Dr. Timothy F. Brady. The dissertation author was the primary author of this paper.

References

Bainbridge, W. A., Isola, P., & Oliva, A. (2013). The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General*, *142*. https://doi.org/10.1037/a0033872

Bartlett, F. C. (1932). Remembering: An experimental and social study. *Cambridge University Press*. (n.d.).

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis*, *20*. https://doi.org/10.1093/pan/mpr057

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, *14*. https://doi.org/10.1016/0010-0285(82)90007-X

Bilalić, M., Langner, R., Ulrich, R., & Grodd, W. (2011). Many faces of expertise: Fusiform face area in chess experts and novices. *Journal of Neuroscience*, *31*. https://doi.org/10.1523/JNEUROSCI.5727-10.2011

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory: Ensemble statistics bias memory for individual items. *Psychological Science*, *22*. https://doi.org/10.1177/0956797610397956

Brady, T. F., Alvarez, G., & Störmer, V. (2019). The role of meaning in visual memory: Face-selective brain activity predicts memory for ambiguous face stimuli. *Journal of Neuroscience*, *39*. https://doi.org/10.1523/JNEUROSCI.1693-18.2018

Brewer, W. F., & Treyens, J. C. (1981). Role of schemata in memory for places. *Cognitive Psychology*, *13*. https://doi.org/10.1016/0010-0285(81)90008-6

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*. https://doi.org/10.1177/1745691610393980

Calkins, M. W. (1894). Experimental. *Psychological Review*, *1*. https://doi.org/10.1037/h0065852

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*. https://doi.org/10.1016/0010-0285(73)90004-2

Curby, K. M., Glazek, K., & Gauthier, I. (2009). A visual short-term memory advantage for objects of expertise. *Journal of Experimental Psychology: Human Perception and Performance*, *35*. https://doi.org/10.1037/0096-1523.35.1.94

De Groot, A. D. (1946). Het denken van den schaker: Een experimenteel-psychologische studie *[The thinking of the chess player: An experimental-psychological study]. Noord-Hollandsche Uitgevers Maatschappij.* (n.d.).

Difallah, D., Filatova, E., & Ipeirotis, P. (2018). Demographics and dynamics of Mechanical Turk workers. *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (pp. 135–143). ACM. https://doi.org/10.1145/3159652.3159661*. (n.d.).

D'Orsi, C. J., Getty, D. J., Pickett, R. M., Sechopoulos, I., Newell, M. S., Gundry, K. R., Bates, S. R., Nishikawa, R. M., Sickles, E. A., Karellas, A., & D'Orsi, E. M. (2013). Stereoscopic digital mammography: Improved specificity and reduced rate of recall in a prospective clinical trial. *Radiology*, *266*. https://doi.org/10.1148/radiol.12120382

Dougal, S., & Rotello, C. M. (2007). "Remembering" emotional words is based on response bias, not recollection. *Psychonomic Bulletin & Review*, *14*. https://doi.org/10.3758/BF03194083

Egan, J. P. (1958). Recognition memory and the operating characteristic. *USAF Operational Applications Laboratory Technical Note, 58–51, ii, 32.* (n.d.).

Engle, R. W., & Bukstel, L. (1978). Memory processes among bridge players of differing expertise. *The American Journal of Psychology*, *91*. https://doi.org/10.2307/1421515

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review*, *102*. https://doi.org/10.1037/0033-295X.102.2.211

Evans, K. K., Cohen, M. A., Tambouret, R., Horowitz, T., Kreindel, E., & Wolfe, J. M. (2011). Does visual expertise improve visual recognition memory? *Attention, Perception, & Psychophysics*, *73*. https://doi.org/10.3758/s13414-010-0022-5

Evans, K. K., Georgian-Smith, D., Tambouret, R., Birdwell, R. L., & Wolfe, J. M. (2013). The gist of the abnormal: Above-chance medical decision making in the blink of an eye. *Psychonomic Bulletin & Review*, *20*. https://doi.org/10.3758/s13423-013-0459-3

Evans, K. K., Haygood, T. M., Cooper, J., Culpan, A.-. M., & Wolfe, J. M. (2016). A half-second glimpse often lets radiologists identify breast cancer cases even when viewing the mammogram of the opposite breast. *Proceedings of the National Academy of Sciences of the United States of America*, *113*. https://doi.org/10.1073/pnas.1606187113

Friedman, A. (1979). Framing pictures: The role of knowledge in automatized encoding and memory for gist. *Journal of Experimental Psychology: General*, *108*. https://doi.org/10.1037/0096-3445.108.3.316

Gauthier, I., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000). Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience*, *3*. https://doi.org/10.1038/72140

Gauthier, I., Tarr, M. J., Anderson, A. W., Skudlarski, P., & Gore, J. C. (1999). Activation of the middle fusiform 'face area' increases with expertise in recognizing novel objects. *Nature Neuroscience*, *2*. https://doi.org/10.1038/9224

Gobet, F., & Simon, H. A. (1996). Recall of random and distorted positions: Implications for the theory of expertise. *Memory & Cognition*, *24*. https://doi.org/10.3758/BF03200937

Graesser, A. C., & Nakamura, G. V. (1982). The impact of a schema on comprehension and memory. In G. H. Bower (Ed.), The psychology of learning and motivation (Vol. 16, pp. 59–109). Academic. (n.d.).

Greene, M. R., Baldassano, C., Esteva, A., Beck, D. M., & Fei-Fei, L. (2016). Visual scenes are categorized by function. *Journal of Experimental Psychology: General*, *145*. https://doi.org/10.1037/xge0000129

Hardesty, L. A., Ganott, M. A., Hakim, C. M., Cohen, C. S., Clearfield, R. J., & Guret, D. (2005). "Memory effect" in observer performance studies of mammograms. *Academic Radiology*, *12*. https://doi.org/10.1016/j.acra.2004.11.026

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, *93*. https://doi.org/10.1037/0033-295X.93.4.411

Hollingworth, A., & Henderson, J. M. (2003). Testing a conceptual locus. *Memory & Cognition*, *31*. https://doi.org/10.3758/BF03196446

Hunt, R. R. (2006). The concept of distinctiveness in memory research. In R. R. Hunt & J. B. Worthen (Eds.), Distinctiveness and memory (pp. 3–25). *Oxford University Press. Https://doi.org/10.1093/acprof:oso/9780195169669.003.0001*. (n.d.).

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, *452*. https://doi.org/10.1038/nature06713

Lampinen, J. M., Copeland, S. M., & Neuschatz, J. S. (2001). Recollections of things schematic: Room schemas revisited. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*. https://doi.org/10.1037/0278-7393.27.5.1211

Light, L. L., Kayra-Stuart, F., & Hollander, S. (1979). Recognition memory for typical and unusual faces. *Journal of Experimental Psychology: Human Learning and Memory*, *5*. https://doi.org/10.1037/0278-7393.5.3.212

McDaniel, M. A., & Einstein, G. O. (1986). Bizarre imagery as an effective memory aid: The importance of distinctiveness. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*. https://doi.org/10.1037/0278-7393.12.1.54

McWeeny, K. H., Young, A. W., Hay, D. C., & Ellis, A. W. (1987). Putting names to faces. *British Journal of Psychology*, *78*. https://doi.org/10.1111/j.2044-8295.1987.tb02235.x

Mickes, L., Flowe, H. D., & Wixted, J. T. (2012). Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous and sequential lineups. *Journal of Experimental Psychology: Applied*, *18*. https://doi.org/10.1037/a0030609

Nosofsky, R. M. (1991). Typicality in logically defined categories: Exemplar-similarity versus rule instantiation. *Memory & Cognition*, *19*. https://doi.org/10.3758/BF03197110

Owens, J., Bower, G. H., & Black, J. B. (1979). The “soap opera” effect in story recall. *Memory & Cognition*, *7*. https://doi.org/10.3758/BF03197537

Pedzek, K., Whetstone, T., Reynolds, K., Askari, N., & Dougherty, T. (1989). Memory for real-world scenes: The role of consistency with schema expectations. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *15*. https://doi.org/10.1037/0278-7393.15.4.587

Rawson, K. A., & Overschelde, J. P. (2008). How does knowledge promote memory? The distinctiveness theory of skilled memory. *Journal of Memory and Language*, *58*. https://doi.org/10.1016/j.jml.2007.08.004

Richler, J. J., Wong, Y. K., & Gauthier, I. (2011). Perceptual expertise as a shift from strategic interference to automatic holistic processing. *Current Directions in Psychological Science*, *20*. https://doi.org/10.1177/0963721411402472

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—Retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*. https://doi.org/10.3758/BF03209391

Vincente, K. J., & Wang, J. H. (1998). An ecological theory of expertise effects in memory recall. *Psychological Review*, *105*. https://doi.org/10.1037/0033-295X.105.1.33

Voss, J. F., Vesonder, G. T., & Spilich, G. J. (1980). Text generation and recall by high-knowledge and low-knowledge individuals. *Journal of Verbal Learning and Verbal Behavior*, *19*. https://doi.org/10.1016/S0022-5371(80)90343-6

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, *19*. https://doi.org/10.1111/j.1467-9280.2008.02136.x

Watson, T. L., & Robbins, R. A. (2014). The nature of holistic processing in face and object recognition: *Current opinions. Frontiers in Psychology, 5*. Https://doi.org/10.3389/fpsyg.2014.00003. (n.d.).

Wells, G. L., Steblay, N. K., & Dysart, J. E. (2011). A test of the simultaneous vs. Sequential lineup methods an initial report of the AJS National Eyewitness Identification Field Studies.https://mn.gov/law-library-stat/archive/urlarchive/a100499.pdf. (n.d.). https://mn.gov/law-library-stat/archive/urlarchive/a100499.pdf

Wixted, J. T. (2005). A theory about why we forget what we once knew. *Current Directions in Psychological Science*, *14*. https://doi.org/10.1111/j.0963-7214.2005.00324.x

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review*, *114*. https://doi.org/10.1037/0033-295X.114.1.152

Wixted, J. T., & Mickes, L. (2015). ROC analysis measures objective discriminability for any eyewitness identification procedure. *Journal of Applied Research in Memory and Cognition*, *4*. https://doi.org/10.1016/j.jarmac.2015.08.007

CHAPTER 2: Visual Hindsight Bias for Abnormal Mammograms in Expert Radiologists.

Abstract

Hindsight bias—where people falsely believe they can accurately predict something once they know about it—is a pervasive decision-making phenomenon, including in the interpretation of radiological images. Evidence suggests it is not only a decision-making phenome- non but also a visual perception one, where prior information about an image enhances our visual perception of the contents of that image. The current experiment investigates to what extent expert radiologists perceive mammograms with visual abnormalities differently when they know what the abnormality is (a visual hindsight bias), above and beyond being biased at a decision level. N = 40 experienced mammography readers were presented with a series of unilateral abnormal mammograms. After each case, they were asked to rate their confidence on a 6-point scale that ranged from confident mass to confident calcification. We used the random image structure evolution method, where the images repeated in an unpredictable order and with varied noise, to ensure any biases were visual, not cognitive. Radiologists who first saw an original image with no noise were more accurate in the max noise level condition [area under the curve (AUC) = 0.60] than those who first saw the degraded images (AUC = 0.55; difference: p = 0.005), suggesting that radiologists' visual perception of medical images is enhanced by prior visual experience with the abnormality. Overall, these results provide evidence that expert radiologists experience not only decision level but also visual hindsight bias, and have potential implications for negligence lawsuits.

Introduction

Hindsight bias is the tendency to misinterpret original convictions given new evidence (leading to the popular phrase, "hindsight is 20/20"). Sometimes referred to as the "I knew it all along," effect, hindsight bias is a well-studied and robust psychological decision-making phenomenon, whereby people who know the outcome of an event both believe that they could have accurately predicted that outcome, when in fact they could not have, and are also unaware that they are biased by their additional knowledge (Wood, 1978; Fischhoff, 1974; Roese & Vohs, 2012; Guibault et al., 2004). Highlighting the robustness of this effect, Fischhoff (1974) found that participants in a study who received prior information about an event happening, relative to those that did not, "had roughly doubled the perceived odds that [the event] was going to occur."

Visual hindsight bias, the "we saw it all along" effect, is a perceptual subtype of hindsight bias in which prior information about an image enhances our perception of that image (Harley, Carlsen, & Loftus, 2004). This perceptual bias is often studied by presenting participants in an experiment with blurry (noisy) images that slowly resolve into clear images and vice versa (Harley, Carlsen, & Loftus, 2004; Bruner & Potter, 1964; Bernstein et al., 2004). Participants who start with the clear image have more information than those who start with the blurry image, allowing experimenters to test the effects of this knowledge on perception.

Using a version of this technique, Harley, Carlsen, & Loftus (2004) showed that individuals tend to underestimate the influence that visual hindsight bias has on their own perception. Participants were asked to identify at which point they recognized the face of a celebrity, which started out blurry (non-recognizable) and slowly dissolved into the clear, original image (Figure 2.1). When participants were subsequently asked to indicate the level of blurriness at which they themselves first recognized the celebrity's face, they consistently overestimated the degree of blur at which they previously recognized the celebrity—thinking they originally recognized the image when it was blurrier than they actually did. A similar study looked at how visual hindsight bias progresses from childhood through adulthood (Bernstein et al., 2004). They found that once children and adults know the identity of a blurry object, they consistently overestimate their peers' ability to recognize the same blurry object. This expands the findings of the previous study and suggests that visual hindsight bias not only affects our own perception, but also our view of others' perception as well, in addition to being present across the lifespan.

Visual hindsight bias can be elicited in more controlled conditions that do not allow for the effects to arise solely from hysteresis as well. For example, using priming to bring an object to mind is sufficient to allow people to recognize images they would not otherwise, even when they images are presented in a random order so that participants cannot simply "hold" onto their previous interpretation (Sadr & Sinha, 2004). To show this, the authors introduced the random image structure evolution (RISE) method of object distortion, where stimuli are systematically transformed by emerging from noise and then dissolving back into noise (Sadr & Sinha, 2004). Right before viewing an object in some level of noise, participants were presented with a word that either matched the following objects' name, or did not match the objects' name (i.e., a completely unrelated word). The authors found that primed images that matched the prior word were recognized more easily than images that were primed by an unrelated word at matched noise levels. Furthermore, this enhanced recognition occurred even when stimuli were intermixed, where it could not arise from a decision level bias. A similar study showed participants various celebrity's faces transforming from blurry to clear or vice versa bias (Bernstein & Harley, 2007). They found that priming participants with the celebrity's name beforehand increased the effects of visual hindsight bias (Bernstein & Harley, 2007). Taken together, these studies provide further evidence that knowledge of the identity of the blurred or distorted image (whether from seeing the clear image, hearing a sound, or simply being told what it is) leads to enhanced perception of the image compared to images that were not preceded by relevant prior information.

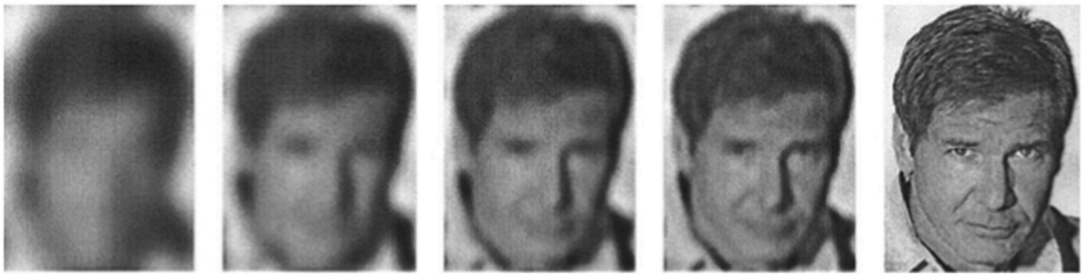


Figure 2.1. Example stimuli used in Harley, Carlsen, & Loftus (2004). This paper provided a clear demonstration of hindsight bias in visual perception. In this example, knowing that the images are of Harrison Ford biases the viewer to recognize Harrison Ford on, say, the second image from the right, even though if you were shown the second image on the right by itself without already knowing its identity, it would be too blurry to recognize.

Visual Hindsight Bias in Experts: Hindsight bias—of the more cognitive variety—has been found in many applied settings and in experts, including in medicine (Muhm et al., 1983), gambling (Toneatto, 1999), legal decision making (Wells & Bradfield, 1998), baseball (Knoll & Srkes, 2017), public policy (Schuett & Wagner, 2011), consumer satisfaction (Zwick, Pieters, & Baumgartner, 1995), and terrorist attacks (Fischhoff et al., 2005). In professional gambling, for example, expert gamblers often reframe losses in hindsight as an event which in retrospect could have been avoided, or reframe wins as confirmation of skill or ability (Toneatto, 1999). Looking at how this bias relates to eyewitness testimony, Wells & Bradfield (1998) showed participants a video of a crime and later asked the participants to identify the suspect. After identification, feedback was given to either confirm or disconfirm their choice. The authors found that confirming versus disconfirming the eyewitnesses' choice had a significant impact on many judgment reports, including the eyewitnesses' self-assessment of their visual experience of the perpetrator (e.g., view, ability to make out facial features, and ease of making identification) (Wells & Bradfield, 1998). These results suggest that the eyewitness is unable to accurately recall the witnessing experience because of this retrospective information. Wixted, Mickes, & Fisher (2018) and others have argued that this contamination of eyewitness memory has caused the prevailing view of the unreliability of eyewitness testimony and suggest that the original judgment, without feedback, is a reliable source of information—but simply one that can be easily corrupted by hindsight biases, new memory encoding, and more, after the initial identification has occurred.

Medical experts are also not immune to cognitive forms of hindsight bias. In one study, neuropsychologists were asked to estimate the probability of three different diagnoses (Arkes et al., 1981). Half of the participants, labeled the hindsight group, were told one of the three diagnoses was correct. The other half of participants who did not receive this "correct" diagnosis were called the foresight group. Of the hindsight group, 58% of participants gave a higher probability estimate than the foresight group to the diagnosis they were told was correct (Arkes et al., 1981).

What about the more visual form of hindsight bias, wherein people report being able to visually see information after they have knowledge of this information from an independent source? In one such study, 82% of cases that had initially been deemed normal by 2 to 3 physicians were later discovered to contain tumors “visible in retrospect,” as far back as 53 months prior to diagnosis (Mumh et al., 1983). Another study looked at visual hindsight bias as it relates to radiologists’ perception of pulmonary nodules (Chen et al., 2020). Radiologists were shown a series of abnormal chest images and asked to either manually add blur until they could no longer see the nodule (hindsight bias condition), or reduce the blur until they could see it (foresight condition). Their results suggest that radiologists are influenced by hindsight bias and that the extent of the bias seemed to be exacerbated with more difficult nodules. While participants report their visual perception, making this a form of visual hindsight bias, blur is manipulated continuously by participants. Thus, unlike the technique of Sadr & Sinha (2004), this result could potentially arise from decision-level biases rather than arising purely as an effect of visual recognition.

The consequences of visual hindsight bias in radiology can be acute. One article describes this anecdotally (Berlin, 2000): when a radiologist looked at an elderly man’s chest x-ray, they concluded that it was normal. The man, however, later became sick and had an additional scan that showed a noticeable mass that eventually led to his death. The man’s family sued the radiologist for initially missing the mass earlier on, when a diagnosis could have prevented the man’s death. In the lawsuit, the case was sent to a second radiologist whose task was to assess and determine whether the mass could have been seen in the original scan. The second radiologist could indeed see the mass (Berlin, 2000). This sequence of events is common in radiology (Berlin & Berlin, 1995); Hugh & Dekker, 2009). When the case is sent to a second radiologist, this physician has additional information when they look at the image in question compared to what the first radiologist had. Depending on the extent of visual hindsight bias in expert radiologists, this additional information could significantly bias their judgment. It could also bias the jury, thus having important legal implications for the radiologist in question. While mammography is one of the most common areas within radiology to be sued for negligence (Baker, 2014), there is relatively limited research on hindsight bias in expert radiologists, and even less research on the effects of visual hindsight bias in radiologists who specialize in mammography. Furthermore, many of the studies of visual hindsight bias allow for a more cognitive interpretation—adding blur in a continuous manner to an image could result in biases because of decision-level hysteresis, for example.

The current study investigates to what extent expert radiologists demonstrate visual hindsight bias to mammograms with visible abnormalities. We use the RISE method (Sadr & Sinha, 2004) to ensure that our results do not arise from decision-level hindsight bias and instead are visual in nature. In order to take into account any response bias that may arise from asking radiologists to distinguish between these two abnormalities, we use receiver operating characteristic analysis to measure performance. To anticipate our results, we find evidence that radiologists are influenced by visual hindsight bias when looking at abnormal mammograms.

Methods

Participants. Aiming for a minimum of 20 participants, we fortunately were able to collect data from 34 radiologists (12 female, 20 male, 2 preferred not to say; age ranged 28 to 69; mean 40) who read an average of 2100 mammograms per year. All participants gave informed consent and were not compensated. The experiment was conducted at the Radiological Society of North America 2019 Conference in Chicago, Illinois, United States.

Informed consent procedures were approved by the Institutional Review Board of the University of California, San Diego.

Stimuli and Procedure: On each trial, radiologists viewed a unilateral mammogram for 3 s. The mammograms subtended ~16 × 20 deg of visual angle at an estimated viewing distance of ~60 cm from the screen. All of the mammograms were abnormal, with half of the mammograms containing a mass and half containing a calcification. All images had verified pathology information and were preclassified by independent radiologists who did not participate in this study.

Each mammogram had five versions of itself with different levels of noise, which were created using a degradation process similar to the RISE procedure developed by Sadr & Sinha (2004). The five copies of each mammogram consisted of: 0% noise, 10% noise, 25% noise, 35% noise, and 45% noise. Figure 2.2 is a visualization of these levels of noise for five example mammograms.

The experimental structure consisted of 8 blocks (30 trials/block), where in each block participants viewed all 5 versions of 6 different mammograms. Blocks were organized by noise level; for example, the six mammograms might each start out presented at their most degraded level (e.g., all six images with 45% noise would be presented sequentially in a random order), and as the trials progressed the images would cycle through the noise levels until all of the images were shown with no noise. Other blocks would start with each image at the 0% noise level (no noise), and then each of the six images would cycle through each noise level to become increasingly degraded. While the noise levels for each block varied systematically—either becoming increasingly noisy or increasingly clear—the six images within each noise level were presented in a random order, ensuring participants could not anticipate whether a given image had a mass or calcification purely based on the sequence itself, reducing cognitive/decision-level biases.

Whether the block started with 0% or 45% noise levels was counterbalanced across blocks and participants. Participants were told when they moved on to the next block. While the images would repeat within a block (for each noise level), no images would repeat across blocks. There were a total of 280 trials across all 8 blocks. The six mammograms chosen for each block were manually categorized based on the structure of the breast outline to decrease the likelihood of participants recognizing repeating images within a block and using this to infer the mass versus calcification judgment.

Immediately after each image was presented, participants were shown a screen containing a 6 point confidence scale ranging from (1) confident this image is a mass to (6) confident this image is a calcification. Using a standard computer mouse, participants were asked to indicate their diagnostic confidence. We used confidence ratings instead of yes/no answers to allow for ROC analysis and separate decision bias from performance. There was no time constraint imposed on responding. After participants indicated their confidence, they clicked a button to move on to the next trial.

Before the experiment began, participants were told that each mammogram would contain either a mass or a calcification and that the task was to rate their confidence on which abnormality they thought was present using a rating scale. They were informed that the images would start out degraded and become more clear, or start out clear and become more degraded and were shown an example of a mammogram depicting the five levels of noise. Participants were not informed that the same image would repeat with a different level of degradation within a batch.

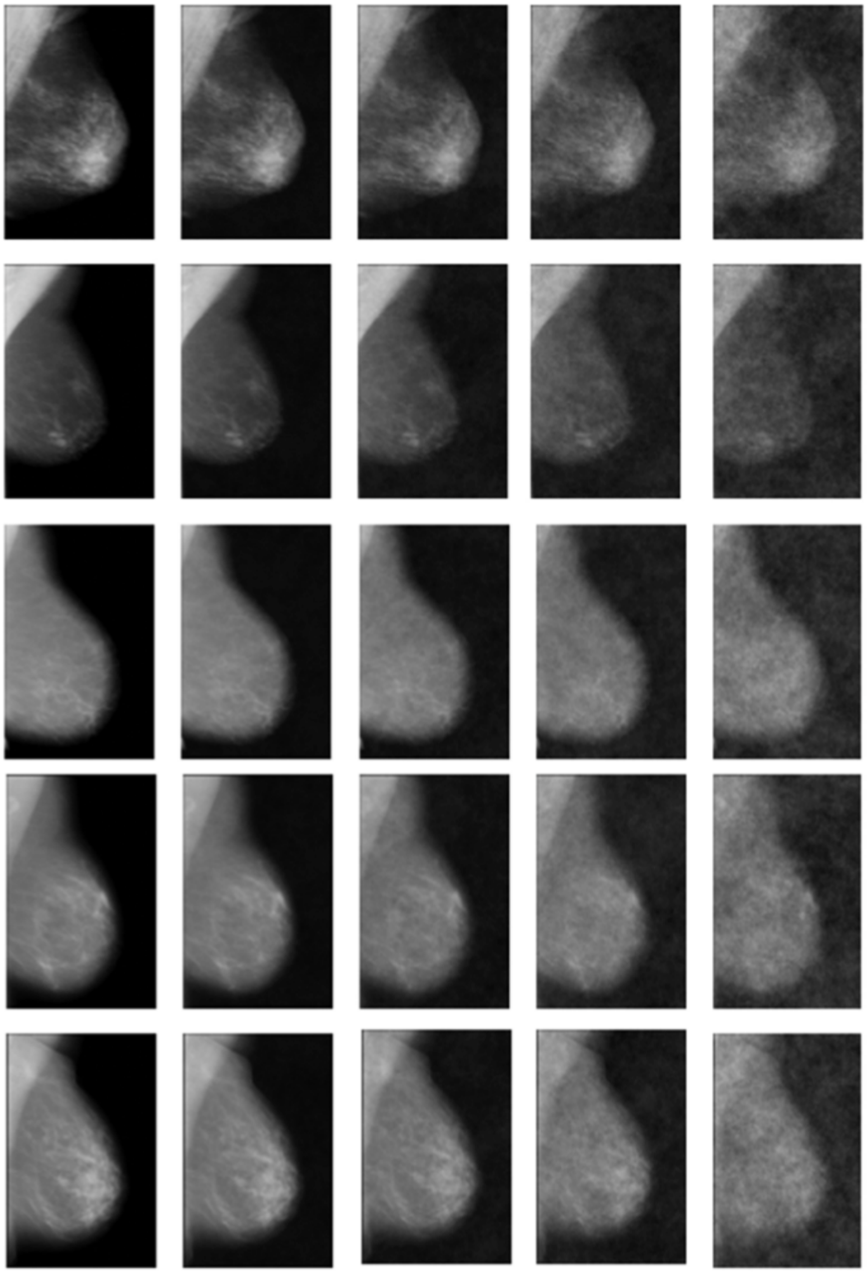


Figure 2.2. Example of the five levels of noise (columns) for five mammogram images (rows).

Analysis: Our main measure was the area under the curve (AUC), an atheoretical measure of discrimination ability. This measure collapses each conditions' ROC to a single measure of performance. Our main hypothesis concerns visual hindsight bias: that is, whether participants might benefit in seeing the abnormality in the noisiest images if they had previously been exposed to the less noisy versions of those images compared to if they had not been exposed to them. In these two situations, participants see the same images, and these are images that are not possible to recognize the abnormality in without previous experience. The highest noise levels provide the best test of hindsight bias because when perceptual information is strong, your "priors" should not play a strong role; however, when perceptual information is weak, visual priors—about how to organize the parts of the image, what objects are where, etc.—will play more of a role. Thus, the highest noise levels provide the core test of visual hindsight bias.

Results

First, we show people's confidence reports (1-6), collapsed across hindsight condition (Figure 2.3, left), with confidence reports shown separately for images that had masses versus images that had calcifications. Next, we show these converted to ROC curves for each noise level (Figure 2.3, right), separated by whether participants started in the hindsight condition (started with 0% noise level) or the no hindsight condition (started with 45% noise level). The ROC is a measure of discrimination performance: The more bowed out the ROC curve is to the top left corner, the better participants were able to complete the task. To quantify these ROCs, the area under the ROC curve, we use AUC, which is an atheoretical measure of overall performance (Figure 2.4).

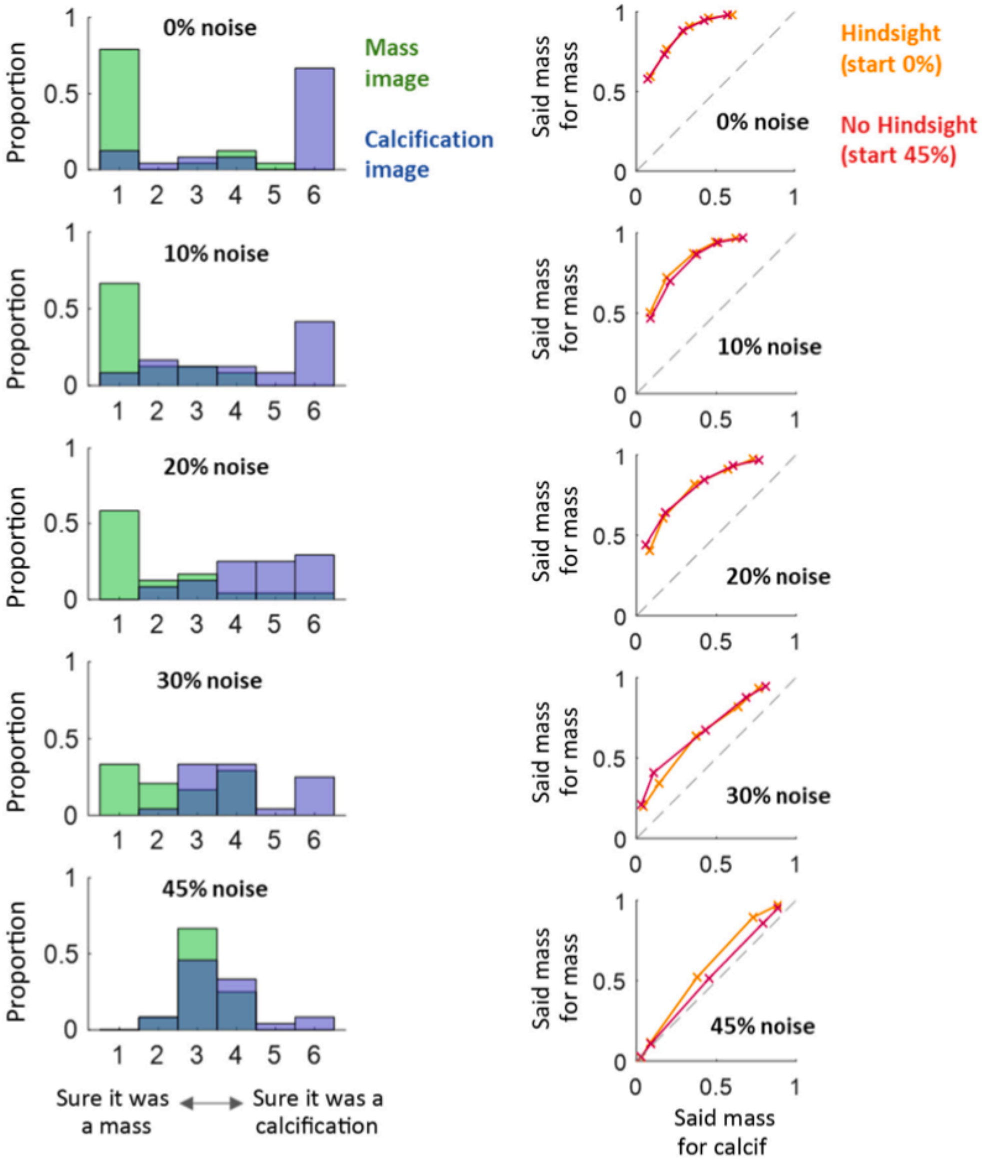


Figure 2.3. Fig. 3 Left: Confidence at different noise levels, collapsing across hindsight conditions. Blue bars show the proportion of responses at each confidence level for calcification images, whereas green bars show the proportion of responses at each confidence level for mass image. Accurate performance is to respond with high numbers if it is a calcification image and low numbers if it is a mass image. As can be seen visually in these data, participants sort the images more accurately when there is less noise. Right: ROCs per hindsight condition per noise level. The confidence data can be converted into an ROC, and then separated by hindsight versus no-hindsight blocks. The more the ROC bows toward the top left corner, the more accurate performance is. We plotted the ROCs in terms of detecting masses, but they are symmetric if you instead plot them in terms of detecting calcification, with no change in the area under the ROC, our measure of interest.

First, considering only how noise affected performance, we found the expected pattern: Collapsing across hindsight conditions, participants were much better at lower noise levels, with performance reliably dropping across noise levels (F(4,132) = 118.3, p < 0.001).

Our main hypothesis concerned visual hindsight bias: that is, whether participants might benefit in seeing the abnormality in the noisiest images if they had previously been exposed to the less noisy versions of those images. To test this, we first did an ANOVA with hindsight condition and noise level as the two factors. We found a main effect of noise (F(4,132) = 117.9, p < 0.001), no main effect of hindsight condition (F(1,33) = 0.50, p = 0.48), but a significant interaction (F(4,132) = 3.25, p = 0.01). This interaction is evidence in favor of our hypothesis that at higher noise levels in particular, there is a benefit to having seen the images previously in the block (hindsight).

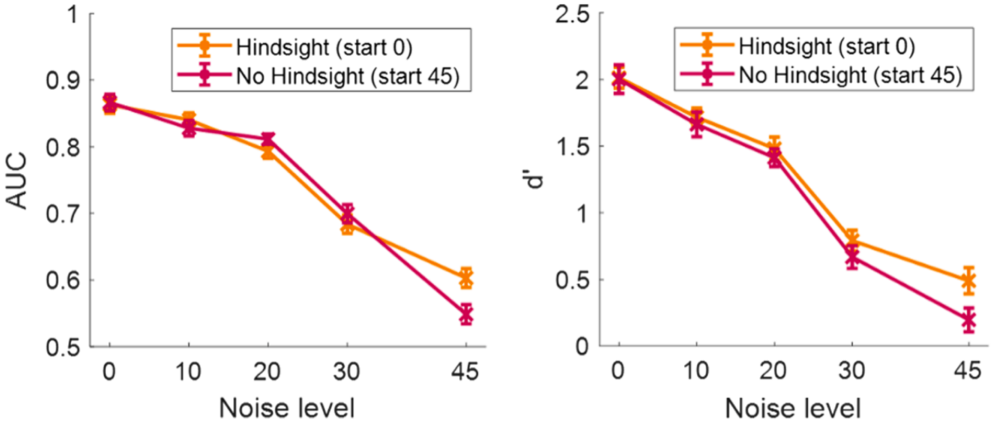


Figure 2.4. AUC (area under the ROC curve) per subject per condition; error bars are within-subject standard error or the mean. An AUC of 0.5 indicates chance performance, and a higher AUC indicates more accurate discrimination of masses and calcifications. Overall, discrimination performance drops with increasing noise, but at the highest noise levels, the hindsight bias condition leads to higher performance.

We also more specifically contrasted performance at the highest noise level between the two hindsight conditions, which was our a priori prediction of where we would expect the largest difference in performance. We found a reliable difference (t(33) = 2.98, p = 0.005, dz = 0.51).

Overall, this data suggest that radiologists who had more information starting out performed better when viewing the most degraded image, compared to radiologists who did not have that prior visual experience: in blocks where radiologists first saw an image with no noise, they did significantly better when the image was maximally noisy, compared to the blocks where they started with a noisy image.

Discussion

The current study found evidence that expert radiologists are influenced by visual hindsight bias when reading mammography images. Our findings support previous research that has shown that hindsight bias is not only a cognitive, decision making bias but also one that affects perception, including expert's perception of medical images. For instance, a recent study showed that radiologists experience visual hindsight bias when looking at pulmonary lung nodules (Chen et al., 2020). The current study expands on these results by providing evidence that expert radiologists who view abnormal mammograms are also not immune to this bias, and using a technique—where images are interleaved—that ensures the results arise from visual perception rather than decision making (Sadr & Sinha, 2004). This is especially pertinent as mammograph radiologists are one of the most commonly sued groups in medicine for negligence (Baker, 2014).

Several potential mechanisms of visual hindsight bias have been proposed. One of the first was increased visual interference. For instance, Bruner & Potter (1964) showed participants a series of common objects that started out of focus and slowly came into focus. Participants who started with very distorted images had more difficulty recognizing the image compared to other groups, showing one of the first experimental instantiations of visual hindsight bias. The authors propose that more distorted visual displays increase the cognitive difficulty of rejecting incorrect hypotheses regarding the identity of the image (i.e., they guess at what it is, incorrectly, hindering later recognition as it gets clearer), whereas clearer images to start allow the observer to better come up with more accurate hypotheses to explain the identity of the image. Later studies support this “creeping determinism,” whereby upon receiving outcome knowledge, the subject immediately integrates this new knowledge with what is already known (Fischhoff, 1974). By testing visual hindsight bias at the highest noise level, we find evidence that when the perceptual input or signal is weak, available prior information—in this case, correct information based on a less noisy version of the image—is integrated into the radiologists knowledge, which allows them to come up with and rely on more accurate hypotheses about the image.

Previous studies have suggested that the strength of hindsight bias varies depending on the difficulty of the perceived information (Chen et al., 2020; Hawkins & Hastie, 1990; Gray, Beilock, & Carr, 2007). For instance, one study found that radiologists had greater visual hindsight bias for more difficult lung nodule cases (Chen et al., 2020). Future research should analyze whether the extent of hindsight bias differs in radiologists depending on the lesion type (e.g., masses, calcifications, and architectural distortions), in addition to difficulty level. Masses and calcifications have very different visual properties (i.e., they vary in size, shape, contrast etc.), which might alter their respective influence on a perceptual bias. Because the task in our study was to indicate whether each image contained a mass or calcification, our study was not designed to address this question: radiologists being prone to say “mass” more often, or “calcification” more often (a response bias) is indistinguishable from lower versus higher difficulty of the two kinds of abnormality in our data. Additionally, our results do not speak to whether radiologists can detect the abnormality better with hindsight (which would be about whether they can distinguish normal versus abnormal), but only whether they can identify particular characteristics of it (mass versus calcification).

It is also unclear whether hindsight bias, whether cognitive or visual, is greater in experts than non-experts within their domain of expertise. Knoll & Arkes (2017) found that expertise exacerbated the bias, whereby baseball experts exhibited systematically greater hindsight bias as the level of the expertise in baseball rules and terminology increased. The authors attributed this effect to “feeling-of-knowing,” which they suggest arises only when expertise is acquired. Other studies have come to a similar conclusion, suggesting that the greater amount of relative knowledge accessible to experts results in an increase in hindsight bias (Musch & Wagner (2007). This is similar to a type of error known as Goldovsky errors, which are known to arise only with expertise. Other studies have shown that experts are less likely to experience hindsight bias (Gray, Beilock, & Carr, 2007; Calvillo & Rutchick, 2014). Calvillo & Rutchick, 2014 show that political expertise was negatively correlated with hindsight bias of predictions made for the 2012 election. Other studies have shown no relationship between expertise and hindsight bias, or suggests that it depends on hypothetical versus actual predictions (Guibault et al., 2004; Calvillo & Rutchick, 2014). Roese & Vohs (2012) attributes many of these discrepancies in the literature to differential mechanisms that either reduce or increase hindsight bias in expertise. To speak towards this ambiguity in the literature, future studies should assess how hindsight bias develops as novices gain experience in their field of expertise. In the current work, naive participants would be unable to accurately perform a mass versus calcification task at all without significant training, and so our task—which was designed solely for radiologists— cannot address this question.

The literature on whether hindsight bias can be reduced is mixed. As suggested by Chen et al. (2020), warning radiologists of the effects of hindsight bias before being presented with the same images they saw earlier may decrease hindsight bias effects. Whether the perceptual bias was reduced or radiologists were adjusting their response to match what they thought the desired outcome was remains unclear. Alternatively, Harley, Carlsen, & Loftus (2004) found that warning participants about hindsight bias did not mitigate the effects of hindsight bias when viewing faces at varying degrees of distortion. Anderson et al. (2014) showed that attempts to eliminate or reduce hindsight bias in judges had no significant effect. The difference in effects across these studies could be attributed to differences in stimuli and experimental design. Given the importance of this bias to many applied fields, future research could expand the current literature on ways to reduce or eliminate this bias with a focus on expert populations. Addressing the malpractice lawsuits specifically, future research could contribute to an emerging field that looks at hindsight bias mitigation strategies for juries (Conklin, 2021).

In summary, this study has provided evidence that expert radiologists are influenced by visual hindsight bias for abnormal mammograms. Future research could investigate whether and how visual hindsight bias changes as novices become experts and whether different categories of abnormalities have an impact on the strength of the bias. The answers to these questions will both expand the current literatures on perceptual biases and expertise, as well as have practical applications in the event a radiologist is sued for negligence.

Chapter 2, in full, is a reprint of the material as it appears in the Journal of Medical Imaging, 2023, co-authored with Samantha Gray and Dr. Timothy F. Brady. The dissertation author was the primary author of this paper.

References

Anderson, J., Jennings, M., Lowe, D., & Reckers, P. (1997). The mitigation of hindsight bias in judges' evaluation of auditor decisions. *Auditing: A J. Pract. Theory*, *16*, 20–39.

Arkes, H. R., Wortmann, R. L., Saville, P. D., & Harkness, A. R. (1981). Hindsight bias among physicians weighing the likelihood of diagnoses. *Journal of Applied Psychology*, *66*(2), 252–254. https://doi.org/10.1037/0021-9010.66.2.252

Baker, S. R. (2014). U.S. medical malpractice: Some data-driven facts. *Springer International Publishing: Notes of a Radiology Watcher*, 177–179.

Berlin, L. (2000). Hindsight Bias. *American Journal of Roentgenology*, *175*(3), 597–601. https://doi.org/10.2214/ajr.175.3.1750597

Berlin, L., & Berlin, J. W. (1995). Malpractice and radiologists in Cook County, IL: Trends in 20 years of litigation. *American Journal of Roentgenology*, *165*(4), 781–788. https://doi.org/10.2214/ajr.165.4.7676967

Bernstein, D. M., Atance, C., Loftus, G. R., & Meltzoff, A. (2004). We Saw It All Along: Visual Hindsight Bias in Children and Adults. *Psychological Science*, *15*(4), 264–267. https://doi.org/10.1111/j.0963-7214.2004.00663.x

Bernstein, D. M., & Harley, E. M. (2007). Fluency misattribution and visual hindsight bias. *Memory*, *15*(5), 548–560. https://doi.org/10.1080/09658210701390701

Bruner, J. S., & Potter, M. C. (1964). Interference in Visual Recognition. *Science*, *144*(3617), 424–425. https://doi.org/10.1126/science.144.3617.424

Calvillo, D. P., & Rutchick, A. M. (2014). Domain Knowledge and Hindsight Bias among Poker Players: Hindsight Bias in Poker. *Journal of Behavioral Decision Making*, *27*(3), 259–267. https://doi.org/10.1002/bdm.1799

Chen, J., Littlefair, S., Bourne, R., & Reed, W. M. (2020). The Effect of Visual Hindsight Bias on Radiologist Perception. *Academic Radiology*, *27*(7), 977–984. https://doi.org/10.1016/j.acra.2019.09.032

Conklin, M. (2021). I Knew It All Along: The Promising Effectiveness of a Pre-Jury Instruction at Mitigating Hindsight Bias. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3962419

Fischhoff, B. (1974). *Hindsight: Thinking backward*. https://apps.dtic.mil/sti/pdfs/ADA001807.pdf

Fischhoff, B., Gonzalez, R. M., Lerner, J. S., & Small, D. A. (2005). Evolving Judgments of Terror Risks: Foresight, Hindsight, and Emotion. *Journal of Experimental Psychology: Applied*, *11*(2), 124–139. https://doi.org/10.1037/1076-898X.11.2.124

Gray, R., Beilock, S. L., & Carr, T. H. (2007). "As soon as the bat met the ball, I knew it was gone": Outcome prediction, hindsight bias, and the representation and control of action in expert and novice baseball players. *Psychonomic Bulletin & Review*, *14*(4), 669–675. https://doi.org/10.3758/BF03196819

Guilbault, R. L., Bryant, F. B., Brockway, J. H., & Posavac, E. J. (2004). A Meta-Analysis of Research on Hindsight Bias. *Basic and Applied Social Psychology*, *26*(2–3), 103–117. https://doi.org/10.1080/01973533.2004.9646399

Harley, E. M., Carlsen, K. A., & Loftus, G. R. (2004). The "Saw-It-All-Along" Effect: Demonstrations of Visual Hindsight Bias. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(5), 960–968. https://doi.org/10.1037/0278-7393.30.5.960

Hawkins, S. A., & Hastie, R. (1990). Hindsight: Biased judgments of past events after the outcomes are known. *Psychological Bulletin*, *107*(3), 311–327. https://doi.org/10.1037/0033-2909.107.3.311

Hugh, T. B., & Dekker, S. W. A. (2009). Hindsight bias and outcome bias in the social construction of medical negligence: A review. *J. Law Med*, *16*(5), 846–857.

Knoll, M. A. Z., & Arkes, H. R. (2017). The Effects of Expertise on the Hindsight Bias: Hindsight Bias and Expertise. *Journal of Behavioral Decision Making*, *30*(2), 389–399. https://doi.org/10.1002/bdm.1950

Muhm, J. R., Miller, W. E., Fontana, R. S., Sanderson, D. R., & Uhlenhopp, M. A. (1983). Lung cancer detected during a screening program using four-month chest radiographs. *Radiology*, *148*(3), 609–615. https://doi.org/10.1148/radiology.148.3.6308709

Musch, J., & Wagner, T. (2007). Did Everybody Know It All Along? A Review of Individual Differences in Hindsight Bias. *Social Cognition*, *25*(1), 64–82. https://doi.org/10.1521/soco.2007.25.1.64

Roese, N. J., & Vohs, K. D. (2012). Hindsight Bias. *Perspectives on Psychological Science*, *7*(5), 411–426. https://doi.org/10.1177/1745691612454303

Sadr, J., & Sinha, P. (2004). Object recognition and Random Image Structure Evolution. *Cognitive Science*, *28*(2), 259–287. https://doi.org/10.1207/s15516709cog2802_7

Schuett, F., & Wagner, A. K. (2011). Hindsight-biased evaluation of political decision makers. *Journal of Public Economics*, *95*(11–12), 1621–1634. https://doi.org/10.1016/j.jpubeco.2011.04.001

Toneatto, T. (1999). Cognitive Psychopathology of Problem Gambling. *Substance Use & Misuse*, *34*(11), 1593–1604. https://doi.org/10.3109/10826089909039417

Wells, G. L., & Bradfield, A. L. (1998). "Good, you identified the suspect": Feedback to eyewitnesses distorts their reports of the witnessing experience. *Journal of Applied Psychology*, *83*(3), 360–376. https://doi.org/10.1037/0021-9010.83.3.360

Wixted, J. T., Mickes, L., & Fisher, R. P. (2018). Rethinking the Reliability of Eyewitness Memory. *Perspectives on Psychological Science*, *13*(3), 324–335. https://doi.org/10.1177/1745691617734878

Wood, G. (1978). The knew-it-all-along effect. *Journal of Experimental Psychology: Human Perception and Performance*, *4*(2), 345–353. https://doi.org/10.1037/0096-1523.4.2.345

Zwick, R., Pieters, R., & Baumgartner, H. (1995). On the Practical Significance of Hindsight Bias: The Case of the Expectancy-Disconfirmation Model of Consumer Satisfaction. *Organizational Behavior and Human Decision Processes*, *64*(1), 103–117. https://doi.org/10.1006/obhd.1995.1093

CHAPTER 3: The Impact of Naturalistic Occlusions on Face Processing in Ensemble Perception

Abstract

The visual system takes advantage of redundancy in the world by forming representations of summary statistics, a phenomenon known as ensemble perception. Ensemble representations are formed for low-level features like orientation and size and high level features such as facial identity and expression. While recent research has shown that the visual system forms intact ensemble representations even when faces are partially occluded via amodal bars, how ensemble perception is impacted with the addition of naturalistic objects such as face masks or sunglasses is largely unknown. To investigate this, we conducted a series of experiments using continuous report tasks in which faces (either varying in identity or expression, i.e., emotion) were partially occluded with a standard surgical mask or sunglasses. Similar to previous research showing that ensemble perception of faces is robust to incomplete information, we found evidence that participants could still extract the average face even when a significant portion of it was occluded with either face masks or sunglasses. We subsequently examined the role of learning and expectation in extracting the average in a group of partially occluded faces by interleaving stimulus sets together such that observers could not easily learn the features of any stimulus set. Performance was worse when participants were not able to learn the features of the set, suggesting not only that the visual system quickly learns the features of the face wheel being used, but that it uses those expectations to help form the ensemble statistic. Overall, our results suggest that the visual system is able to form robust ensemble representations for faces with naturalistic occlusions, but that robustness appears to be supported by learning information about the particular feature set being used.

Introduction

As we move about the world, we are bombarded with far more information than our visual system can process at one time. Yet this fact is not detrimental to our ability to successfully navigate and process information, in part because the visual system has developed tools and heuristics to counter limited capacity systems. Research suggests ensemble perception serves as a catalyst to allow some information to bypass the severe capacity limitations of visual working memory and attentional systems (Haberman & Whitney, 2012; Whitney, Haberman, & Sweeny, 2014; Alvarez, 2011). Ensemble perception takes advantage of redundancy in the world (e.g., blades of grass, faces in a crowd) by computing the statistical average of visual information quickly and with minimal processing power (Haberman, Brady & Alvarez, 2015; Alvarez & Oliva, 2009). Increasing the number of items on the screen, for instance, has a detrimental impact on the number of individual items we can reliably process, but does not affect our ability to extract the mean (Chong & Treisman, 2003; Alvarez, Brady, & Haberman, 2015). The resulting average is often called the ensemble percept or ensemble representation, and is a useful heuristic to guide decisions and behavior without needing to process every individual instance of redundant visual information.

Ensemble representations support our visual percept throughout all levels of visual processing, operating on simple features such as size, orientation, and motion, as well as for more complex features such as facial identity, expression, gender, and ethnicity (Haberman & Alvarez, 2015; Tanaka & Farah, 1993; Haberman & Whitney, 2007; Smith, Cottrell, Gosselin, & Schyns, 2005; Leib et al., 2014; Haberman, Lee, & Whitney, 2015). Even for high level, complex information such as faces, the ensemble percept can be extracted without the need to fully process – or even be aware of – individual items (Wolfe, Kosovicheva, Leib, Wood, & Whitney, 2015). For instance, participants can still extract the average face even when failing to recognize individual changes driving the average (Haberman & Whitney, 2011), or when unable to recognize individual faces due to crowding or noise (Fischer & Whitney, 2011). Given the ubiquitousness of ensemble perception across many types of information and its ability to relay useful information without needing to process all of the information present, the ensemble heuristic is thought of as a fundamental mechanism supporting our visual experience.

Incoming visual information is noisy and often partially intact, such as when objects are partially occluded by another object (e.g., your car partially obstructed by another parked vehicle), or the viewpoint masks relevant features (e.g., a face turned to the side). Given these real world complexities, there is growing work investigating how both perception of individual items as well as ensemble perception is impacted in the presence of incomplete or limited information. Several studies using amodal stimuli, which prompt the visual system to perceive the "whole", despite part of the stimulus being absent or occluded, have found that ensemble representations of both low and high level information maintain precision and accuracy when presented with amodal (i.e., incomplete) information (Haberman & Ulrich, 2019). In particular, Haberman & Ulrich (2019) found evidence that the ensemble representation of both facial identity and expression can still be derived when portions of faces are amodally covered with artificial black bars, preventing viewers from seeing all of the facial features. Interestingly, the researchers did not find evidence that individuals create and use a 'completed' face in their representation of the missing information; in other words, forming the ensemble percept was not contingent on representing the missing information

(Haberman & Ulrich, 2019). The current study expands on this work by investigating how ensemble perception is impacted by types of occluders that occur more naturally in the world – sunglasses and face masks.

Since the COVID-19 pandemic, there has been a significant increase in research on how more ecologically valid or naturalistic occlusions impact perception of individual faces (Lander & Saunders, 2023). Work looking at face masks has found that face masks generally decrease our ability to recall or recognize the identity of a face (Carragher & Hancock, 2020; Freud et al., 2020), although some information such as a sense of familiarity for famous faces remains intact (Carlaw et al., 2022). Carlaw et al. (2022) also found that recognition impairment was greater when the faces were wearing sunglasses compared to masks, which aligns with a large body of research showing the importance of eyes in recognizing faces (McKelvie, 1976; Daview, Ellis, & Shepherd, 1977).

The challenge that such occlusions pose to individual face recognition occurs at least in part because there is less evidence observers can amass in order to make a decision (Carragher et al., 2022; McKelvie, 1976), leading to masked faces recognition relying more heavily on individual features opposed to holistic processing (Freud et al. 2020). In light of this, Carragher et al. (2022) tested whether a training strategy designed to help observers maximize their awareness and use of available features (a strategy known to be used by professional face examiners) helps mitigate the impairment face masks cause on identification and communication. They found that even brief diagnostic feature training (i.e., telling participants to focus on the available information) improved recognition performance significantly (Carragher et al., 2022). This research suggests that for individual faces with naturalistic occlusions, observers learn to take advantage of available featural information to help overcome the decrease in available usable information from the occluder.

In the current series of experiments, we explore whether our ability to extract the ensemble holds for naturalistic occlusions, in a similar way it does for more artificial amodal bars. In Experiment 2, we interleave conditions and stimulus sets together in order to assess the impact of feature learning on ensemble perception, which has been shown to contribute to face processing of individual faces with masks or sunglasses. Finally, we include a single face condition, which serves as a baseline to determine if adding these occluders disrupts ensemble perception above and beyond any impact on perception of a single face. The results of the current experiments add to a growing literature that looks at how face masks impact face perception (Lander & Saunders, 2023), while also contributing to our understanding of ensemble perception broadly. To preview our results, we find that participants are able to derive the ensemble of faces despite added naturalistic occlusions, but that there is some cost and evidence of a significant role of feature learning.

Experiment 1

In Experiment 1, we explored the impact of face masks and sunglasses on the ensemble representation of facial identity and expression. While there is growing research on the impact of occlusions such as face masks on individual face perception (Lander & Saunders, 2023), the current study is one of the first to investigate the influence of naturalistic occlusions on ensemble perception. By examining both face masks and sunglasses, we can also begin to ascertain whether certain features are weighted differently in the process of extracting the ensemble of facial information.

Experiment 1 was broken down into four sub experiments (described more below) testing each of the four conditions separately: identity masks, identity sunglasses, expression masks, and expression sunglasses.

Experiment 1 Methods

115 total participants completed Experiment 1 (39 in 1A, 18 in 1B, 39 in 1C, and 19 in 1D). All participants were University of California San Diego undergraduate students who gave informed consent and were compensated with course credit. Informed consent procedures were approved by the Institutional Review Board of the University of California, San Diego.

Stimuli and design. Participants were presented with a single face or groups of four faces varying in either identity (Experiments 1A and 1B) or expression (Experiments 1C and 1D). Our stimulus sets were the same as those used in the Haberman and Ulrich (2019) from the Harvard Face Database, which consist of 360 linearly interpolated face morphs based on three male faces. Face morphs were separated in either identity or expression "units," corresponding to approximately equal steps in a 360 degree circular stimulus space. For Experiment's 1A and 1C (mask occluders), we made a copy of each face wheel (identity and expression) and added standard white surgical masks onto the faces, which covered the lower half of the face. For Experiments 1B and 1D, we added opaque black sunglasses onto a copy of both wheels. Taken together, Experiment 1 had six face wheels – identity wheel with no occlusion, identity wheel with masks, identity wheel with sunglasses, expression wheel with no occlusion, expression wheel with masks, and expression wheel with sunglasses.

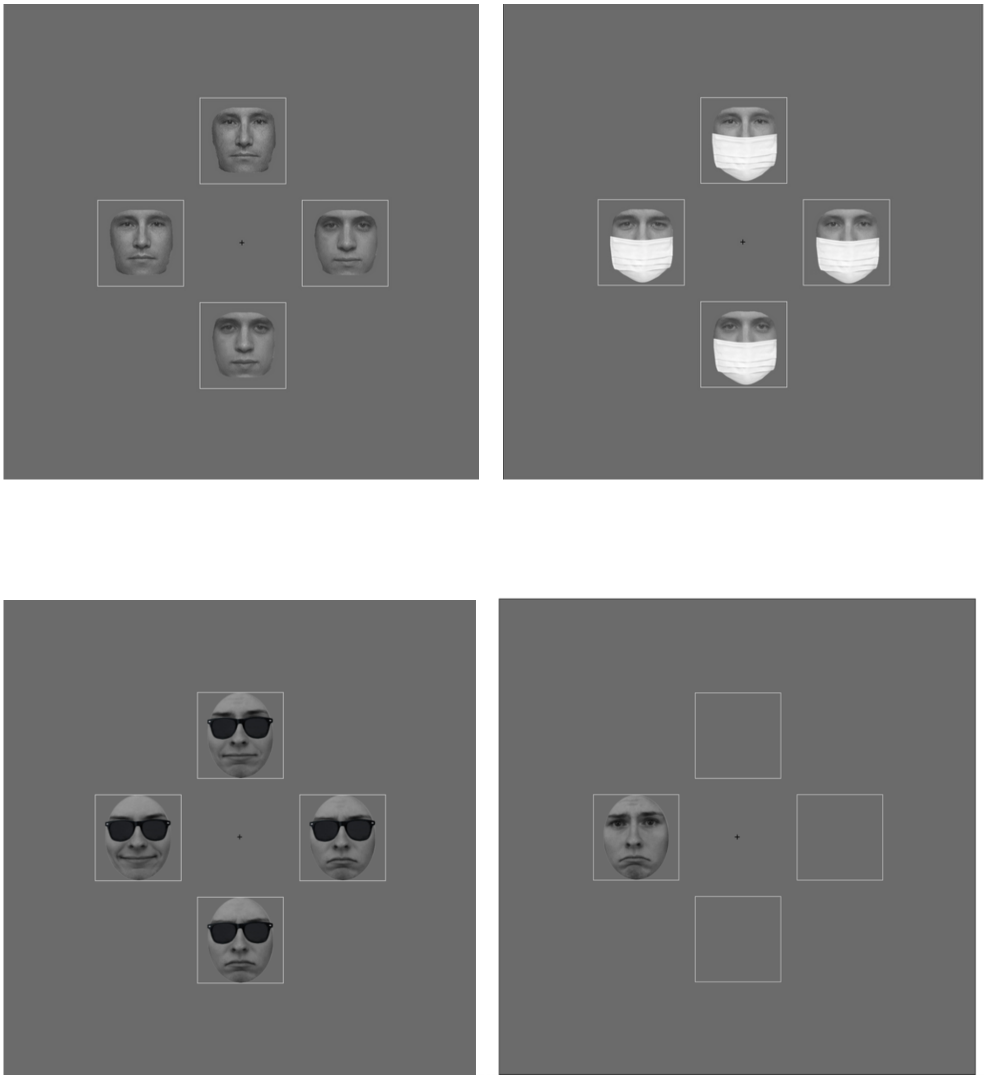


Figure 3.1: A selection of trial types in Experiment 1. The top left shows an unoccluded ensemble identity trial. The top right shows an ensemble identity trial with masks. The bottom left shows an ensemble expression trial with sunglasses. The bottom right shows a single face unoccluded expression trial.

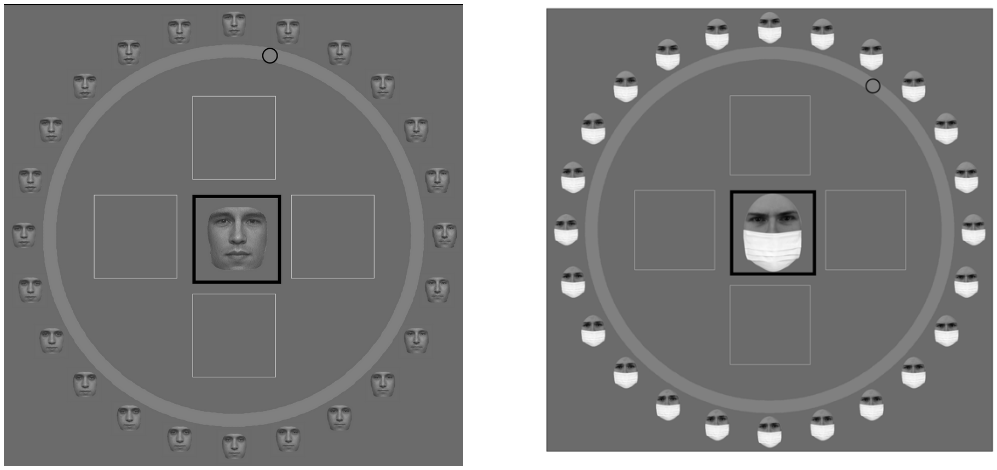


Figure 3.2: A depiction of the test face: participants were told to move their mouse around the wheel until the middle test face matched what they felt represented the ensemble of the previous trial. The test face had the same occlusion as the previous set (i.e., if the face(s) shown in the trial were masked, then the test face was masked).

In each sub-experiment, we tested performance on a single face vs. the ensemble and occlusion vs. no occlusion. This led to each sub-experiment consisting of four counterbalanced blocks: single face (occluded, unoccluded) and ensemble (occluded, unoccluded). The unoccluded conditions serve as a baseline or control to examine the extent to which adding naturalistic occlusions impacts our ability to accurately represent the identity or expression of a single or average of a group of faces. Each face was presented in grayscale and approximately 125 x 125 pixels in size. The mean of each set was chosen randomly on each trial. The spacing of the four faces around the mean is discussed in detail below.

Procedure. The same procedure was followed for experiments 1A, 1B, 1C, and 1D. On each trial in an ensemble block, participants were told to report the average identity or expression of a set of faces. Four faces would appear on the screen for 1 second, followed by a 250-ms interstimulus interval (ISI), and then a single test face appeared on the center of the screen (Figures 3.1 and 3.2). The test face always matched the condition of the previous set (e.g., if four faces with masks were shown, the test face had a mask; if four faces with sunglasses were shown, the test face wore sunglasses, etc.). As participants moved their mouse along the face wheel surrounding the test face (see Figure 3.2), the identity or expression of the test face morphed to match the corresponding location on the wheel. Each degree of change on the circle corresponded to a degree of change in either the identity or expression of the face. Once participants clicked on the location on the wheel that best matched their perceived average of the four preceding faces, they were shown a blank screen and told to click a small black cross in the middle of the screen to initiate the next trial. Performance was measured as the distance in degrees on the circle from where the participant clicked indicating their perceived average to the actual mean of the set.

Single face trials were largely the same as the ensemble trials, except only one face appeared on the screen on each trial. The location of the presented face was randomly chosen to appear in one of the four locations the ensemble set appeared in. Participants were told to replicate their perception of the single face on the test face.

Each of the four blocks contained 80 trials. The total time for each sub-experiment was approximately 40 minutes. Participants were given instructions prior to each block, noting whether they would be viewing a single face or ensemble block, and whether the faces would have masks, sunglasses, or no occlusions.

Experiment 1 Results

Analysis & Data Clustering: In each sub-experiment, error was used as an index of performance, calculated as the circular standard deviation of the distance (in degrees) between the face the participant chose as the average with the true average. Smaller error (i.e., less distance between the true average and the perceived average) indicates a stronger or more precise representation of the ensemble percept or single face. We looked at whether error changed with the addition of naturalistic occluders, both in the single face and ensemble conditions.

When analyzing the results of Experiment 1, we discovered something unintentional about the experimental design: the four faces presented at encoding in ensemble blocks were chosen at varying distances from the mean for each trial. This means that sometimes the four faces presented on a given trial might be clustered closer to the average (an easier trial), maximally separated along the wheel (a nearly impossible trial), or somewhere in between. In typical ensemble perception experiments, group members are all chosen to be an exact number of degrees apart around a random true mean so that difficulty level is controlled for. While unintentional, this organization of the data offered us an opportunity to parse the data by how clustered the faces were on the wheel (i.e., how difficult the ensemble trials were) and examine the impact of clustering distance on ensemble performance. While there has been some research looking at set similarity in ensemble perception for faces (Cabeza, Bruce, Kato, & Oda, 1999), none has assessed this with naturalistic occlusions.

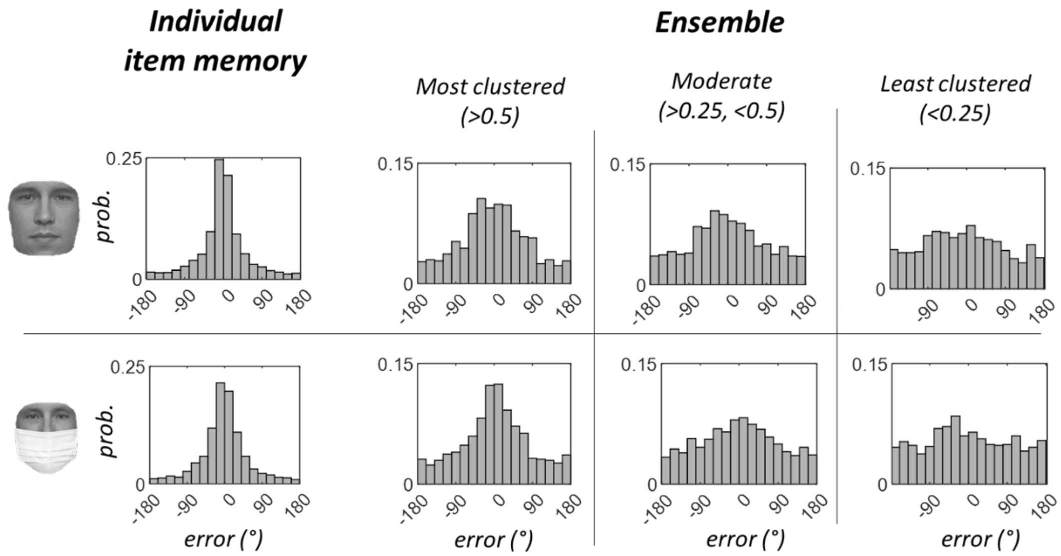


Figure 3.3. Histogram distribution of error from Experiment 1A as an illustration of performance in the individual item condition and the three levels of clustering in the ensemble task. Top row shows non-masked faces and bottom shows masked faces. Errors near 0 are indicative of good performance.

We parsed the data based on the additive vector of the four faces along the wheel. For example, if all four faces were chosen to be the same as the mean (which never occurred, but just as an example), the additive vector would be 1. If the four faces were as far apart as they could be, the vector would be zero. When the four faces are clustered close together on the wheel – such that the additive vector of the four faces along the wheel is 0.8 or greater, for example – we expect error to be reduced. When the four faces are located further apart from the mean, we anticipate the error to be greater as it is almost impossible to choose a spot on the wheel that averages four almost equidistant locations. We parsed the data into thirds, which correspond to vector lengths of 0-0.25, 0.25-0.5, and 0.5-1. See figure 3.3 for a visualization of the error distributions based on the three clusters, using Experiment 1A's data as a case study. In each of the four sub-experiments below, we utilized trials where the ensemble vector was greater than 0.5, which corresponds to a moderate cluster and is reflected by the thickest and lowest red line in Figure 3.4.

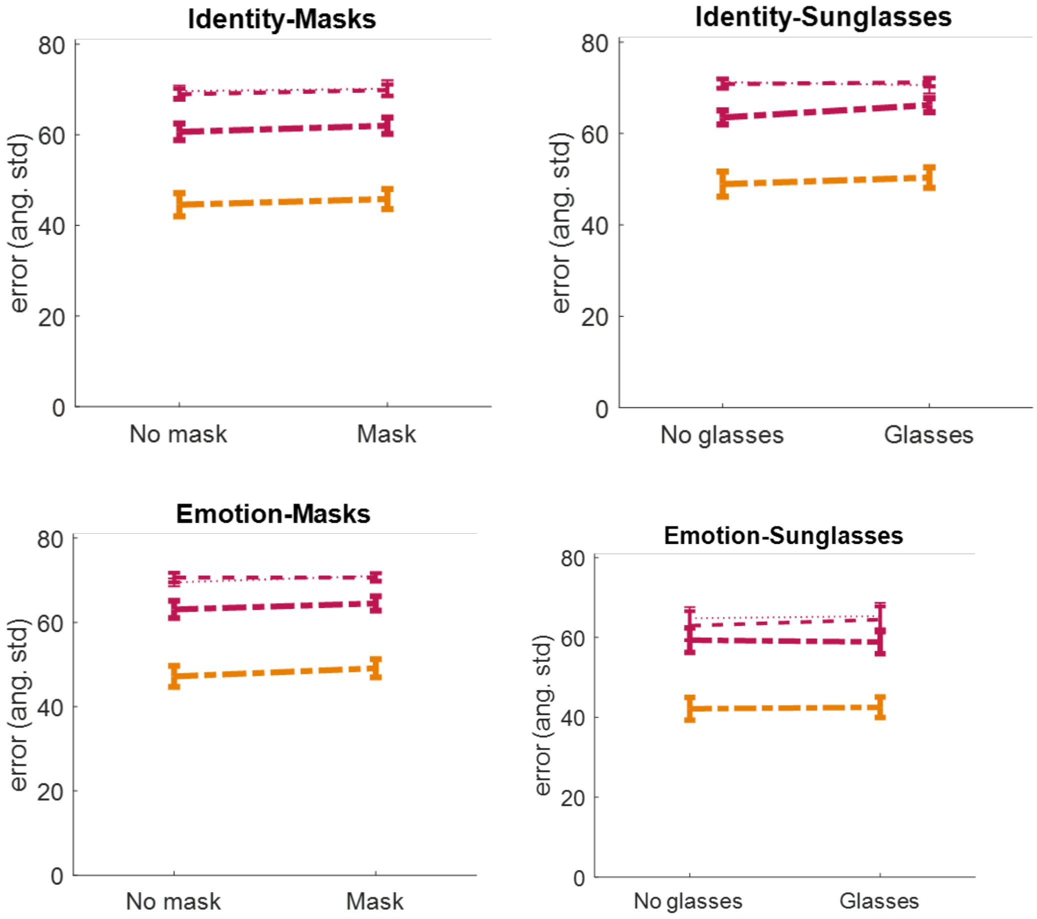


Figure 3.4. Each yellow bar represents performance for the single face trials and the red bars represent performance for the ensemble trials and are broken down by the 3 cluster levels. The thin dotted line corresponds to a vector length of <0.25, medium dashed line corresponds to a vector length of 0.25-0.5, and the thick dashed line corresponds to the vector length of >0.5. **Top left:** The results of Experiment 1A showing the impact of adding a mask on extracting the average identity of a group of faces. **Top right:** Results of Experiment 1B showing the impact of adding sunglasses on extracting the average identity of a group of faces. **Bottom left:** Results of Experiment 1C showing the impact of adding a mask on extracting the average expression of a group of faces. **Bottom right:** Results of Experiment 1D showing the impact of adding sunglasses on extracting the average expression of a group of faces.

Experiment 1A Results - Identity Masks: Experiment 1A examined the impact of masks on ensemble perception for facial identity and followed the procedure described above. This experiment had four blocks, counterbalanced across participants: single face (masked, no mask) and ensemble (masked, no mask).

We first analyzed the impact of adding face masks on our perception of a single face. We found that our ability to accurately represent a single face was not impacted by the addition of a mask (t(38)=-0.76, p=0.453, dz=0.12). Next we asked whether the addition of face masks hurt our ability to extract the average facial identity of the group of four faces varying in identity. We found that our ability to extract the statistical average of facial identity was not impacted by the additional mask (t(38)=-0.78, p=0.440, dz=0.13), which suggests that the ensemble heuristic is robust to the added occlusion.

Experiment 1B Results - Identity Sunglasses: Experiment 1B examined the impact of sunglasses on identity ensemble perception and followed the procedure described above. This experiment had four blocks, counterbalanced across participants: single face (sunglasses, no sunglasses,) and ensemble (sunglasses, no sunglasses).

Similar to Experiment 1A, we first analyzed the impact of adding sunglasses on our perception of a single face. We found that, similar to masks, our ability to accurately represent a single face was not significantly impacted by the addition of sunglasses (t(38)=-0.99, p=0.327, dz=0.16). Next we asked whether the addition of sunglasses hurt our ability to extract the average identity of a set of four faces varying in identity. Similar to Experiment 1A, our ability to extract the statistical average of facial identity was not impacted by the additional pair of sunglasses (t(38)=-1.66, p=0.106, dz=0.27), which provides additional evidence that the ensemble heuristic is robust to the added occlusion.

Experiment 1C Results - Expression Masks: Experiment 1C examined the impact of face masks on expression ensemble perception and followed the procedure described above. This experiment had four blocks, counterbalanced across participants: single face (masked, no mask) and ensemble (masked, no mask).

Our findings were consistent with Experiment 1A and 1B. In analyzing the impact of face masks on our perception of a single face, we found no significant change in performance with the addition of a mask (t(38)=-1.45, p=0.154, dz=0.23). Next we asked whether the addition of face masks hurt our ability to extract the average facial expression of the set. Similar to previous experiments, we found that our ability to extract the statistical average of facial expression was not impacted by the additional mask (t(38)=-1.02, p=0.314, dz=0.16).

Experiment 1D Results - Expression Sunglasses: Experiment 1D examined the impact of sunglasses on expression ensemble perception and followed the procedure described. This experiment had four blocks, counterbalanced across participants: single face (sunglasses, no sunglasses) and ensemble (sunglasses, no sunglasses).

The findings of Experiment 1D were consistent with 1A-1C. We first analyzed the impact of adding sunglasses on our perception of a single face. We found that our ability to accurately represent a single face was not impacted by the addition of a pair of sunglasses (t(38)=-0.18, p=0.860, dz=0.03). Critically, we next asked whether the addition of sunglasses hurt our ability to extract the average facial expression of the set. Similar to the experiments above, we once again found that our ability to extract the statistical average of facial expression was not impacted by the additional pair of sunglasses (t(35)=0.26, p=0.796, dz=0.04). Taken together, these results suggest that ensemble perception is not impacted significantly by the addition of face masks or sunglasses.

Experiment 2: Feature Learning

In Experiment 1, we found evidence that ensemble perception is quite robust: despite a significant drop in available information when faces are occluded, participants can still extract the average identity and expression of the group. However, Experiment 1 does not speak towards any strategies that participants may be using to derive the ensemble representation in these non ideal environments.

Experiment 2 has two goals: to replicate Experiment 1 in blocked conditions using controlled stimulus clustering, and to assess feature learning as a potential strategy used to extract the average despite incomplete facial information. Research looking explicitly at ensemble perception of faces has suggested that missing information is not represented and averaged, or at least not well (Haberman & Ulrich, 2019). In addition, recent work looking at how individual masked faces are processed provide evidence that people focus on the available information (i.e., maximize the evidence gained from the visible features) and rely on that information to make decisions (Carragher et al., 2022). Experiment 2 tests these ideas in ensemble perception using naturalistic occlusions.

Methods

143 participants completed the blocked experiment, and 28 participants completed the interleaved experiment. Participants were University of California San Diego undergraduate students. All participants gave informed consent and were compensated with course credit. Informed consent procedures were approved by the Institutional Review Board of the University of California, San Diego.

Stimuli, design, & procedure: Similar to Experiment 1, participants were presented with sets of faces varying in either emotion or identity and either wearing a mask, sunglasses, or no occlusion. As we primarily wanted to compare blocked vs. interleaved designs to assess the impact of feature learning, we did not include a single face condition in Experiment 2. Similar to Experiment 1, participants completed 2 conditions out of 8 possible conditions in the blocked experiment, such that the only difference in was the controlled clustering of the set of faces: ensemble identity (masked and unmasked, or sunglasses and no sunglasses) and ensemble expression (masked and unmasked, or sunglasses and no sunglasses).

In the interleaved experiment, all of the trials from the blocked experiment were interleaved together, such that participants did not know which type of trial type came next. Each participant in the interleaved experiment completed all six trial types. A participant might have a trial asking them to average four faces with masks varying in identity, followed by four faces wearing sunglasses varying in expression, followed by four faces varying in expression with no occlusion, etc. This design discouraged participants from learning the features of any particular stimulus wheel as they did not know whether that information would be useful on any given trial. We predict that if participants are maximizing their use of knowledge about the available features to create the ensemble percept, then by disrupting their ability to as easily learn the stimulus features will lead to a drop in performance.

The procedure for each trial was largely the same as in Experiment 1, where participants were shown an ensemble of four faces for one second followed by a 250 ms ISI and test face, where participants indicated the best representation of the perceived average of the four previous faces. Unlike in Experiment 1, all of the presented faces varied consistently in -45, -15, +15 and +45 units from the mean, which is in line with previous experiments and corresponds to a circular vector length of 0.84.

Results

The results for Experiment 2 are displayed in Figure 3.5. We first analyzed the data from the blocked experiment, asking whether face masks or sunglasses hurt ensemble performance. Note the only difference in the blocked version of Experiment 2 and Experiment 1 was that the clustering of faces presented was controlled in all conditions. We subsequently analyzed data from the interleaved experiment, and then compared blocked versus interleaved designs.

We first analyzed how the addition of face masks or sunglasses impacted our ability to extract the average expression. We found that both types of occlusions hurt performance significantly, with sunglasses having a greater negative impact on performance (Adding masks: $t(35)=2.25$, $p=0.029$; adding sunglasses: $t(36)=6.10$, $p<0.001$). While these differences are significant, performance is still relatively strong. For instance, a one sample t-test indicates that ensemble expression performance with either masks or sunglasses is still reliably different from chance, which was simulated as 81 degrees (Masks v chance: $t(35)=11.7$, $p <0.001$; sunglasses v chance: $t(36)=13.8$, $p<0.001$).

Next we analyzed whether the added occlusions impacted our ability to extract the average identity in the blocked experiment. Similar to Experiment 1, we found that adding masks did not significantly harm our ability to extract the average identity of a group of faces ($t(31)=1.08$, $p=0.286$). Adding sunglasses, however, did hurt our ability to extract the average identity ($t(37)=4.61$, $p<0.001$). Similar to the expression trials, adding face masks or sunglasses did not impact performance so much so that it was anywhere near chance (masks v chance: t(31)=11.0, p<0.001; sunglasses v chance: t(37)=8.32, p<0.001).

Apart from the impact of masks on extracting the average identity, adding occlusions in all other conditions hurt performance, unlike in Experiment 1. This difference between experiments is likely due to the difference in clustering: by controlling the spacing between the ensemble set members and thus being a more targeted test, Experiment 2 is able to reveal some reliable cost to adding the occlusions. However, this cost is far from chance performance in all conditions, suggesting that ensemble perception is still a powerful heuristic despite being influenced by the disruption in information.

Next, we analyzed the results from the interleaved experiment. All participants completed blocks that had all 8 trial types intermixed together, in order to discourage the ability to learn the particular features of any stimulus wheel. A one-way repeated-measures ANOVA revealed a significant effect of condition (F(1,40)=16.0, p<0.001). We first looked at the impact of occlusions on the ability to extract the average identity. Pairwise *t* tests revealed that both masks and sunglasses had a significantly detrimental effect on ensemble performance for facial identity (adding masks: t(40)=2.91, p=0.006; adding sunglasses: t(40)=5.10, p<0.001. However, similar to the blocked experiments, performance in the occluded conditions were still reliably different from chance (masks v chance: t(40)=9.79, p<0.001; sunglasses v chance: t(40)=10.1, p<0.001).

Next we looked at the impact of occlusions on the ability to extract the average expression. Similar to identity, pairwise *t* tests revealed that both masks and sunglasses had a significantly detrimental effect on ensemble performance for facial expression (adding masks: t(40)=4.85, p<0.001; adding sunglasses: t(40)=4.81, p<0.001. However, similar to the blocked experiments, performance in the occluded conditions were still reliably different from chance (masks v chance: t(40)=10.6, p<0.001; sunglasses v chance: t(40)=11.6, p<0.001).

Taken together, we find a reliable cost to adding the occlusion, but that we are still effectively able to extract the ensemble nonetheless. This leads to our final analysis looking at the difference between blocked and interleaved conditions.

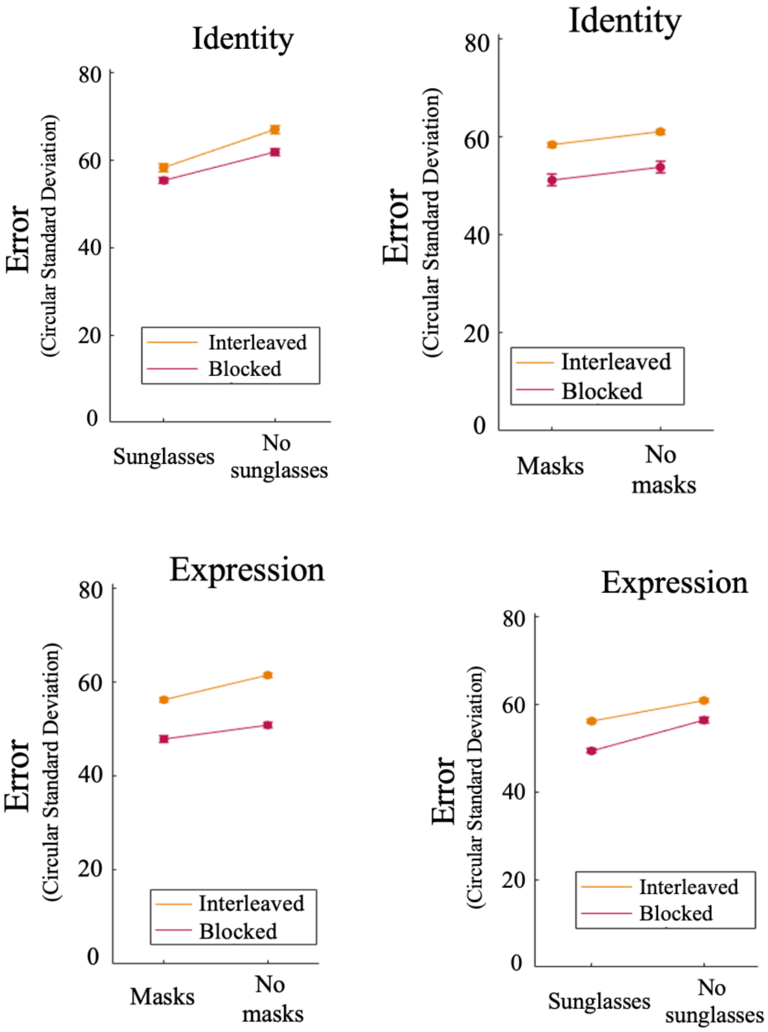


Figure 3.5. Above: Results within the identity wheels. Below: Results from expression face wheels. The yellow bars are the results from the interleaved experiment and the red bars are the results of the blocked experiments.

To compare blocked vs interleaved experiments, we conducted a 2x2 mixed effect ANOVA for each of the four condition pairs. Comparing blocked vs. interleaved conditions for extracting the average identity of faces with or without masks, we found a main effect of masks (F(1,71)=4.94, p=0.029) as well as a main effect of condition (i.e., blocked vs. interleaved) (F(1,71)=5.96, p=0.017). Comparing blocked vs. interleaved conditions for extracting the average identity of faces with or without sunglasses, we found a main effect of sunglasses (F(1.77)=46.36, p<0.001), but no main effect of condition.

Next we assessed the expression blocks. Compared blocked vs. interleaved conditions for extracting the average expression of faces with or without masks, we found a main effect of masks (F(1,75)=23.88, p<0.001) as well as a main effect of condition (F(1,75)=10.6, p=0.002). Comparing blocked vs. interleaved conditions for extracting the average expression of faces with or without sunglasses, we found a main effect of sunglasses (F(1,76)=60.92, p<0.001), as well as a main effect of condition (F(1,76)=4.65, p=0.034).

Overall, we find that one of the biggest effects is that of the interleaved design: subjects had greater error in all conditions compared to the blocked design, which was significant in all but the identity sunglasses condition. These results may suggest, at least in part, that participants are learning feature information about the set in the blocked case and those expectations help them in the blocked conditions.

General Discussion

The present study examines whether the ability to extract the average facial identity or expression is impacted when faces have naturalistic occlusions such as face masks or sunglasses, and what strategies participants might employ to combat the added occlusion. We found that while there is some cost to adding either type of occlusion, it is not enough to push observers to chance performance. Taken together, our results suggest that ensemble perception for faces can still operate despite the obstruction of information.

In Experiment 1, we found the spacing of the individual items in the ensemble set varied around the mean on each trial, such that the individual items of the ensemble might be more or less clustered around the mean. To account for this, we parsed the data into thirds, corresponding to how far apart the items were from the mean. We used the top third cluster (the third of the data corresponding to the most clustered trials) to compare occluded vs unoccluded ensemble performance in our analysis. It is established that ensembles with less variance lead to better performance. For example, ensembles with less color variation makes it easier to extract the average color (Maule & Franklin, 2015), and extracting the mean face is easier when the faces of the set are more similar to the average (Cabeza, Bruce, Kato, & Oda, 1999). Figure 3.3 is a visualization of a similar pattern in our experiment: the more clustered the faces were towards the mean, the more performance tended to be closer to the true average.

Experiment 2 was designed to 1) replicate Experiment 1 with controlled ensemble spacing, and 2) to test whether participants rely on a feature learning strategy to form the ensemble percept in situations with limited available information. In the blocked condition, and unlike in Experiment 1, we found that adding sunglasses or face masks hurt participants' ability to extract the average expression or identity of the set compared to unoccluded ensembles. However, in all cases, the decrease in performance did not push participants to chance, or even close to it. Thus, although there is some cost, the blocked condition in Experiment 2 still provides strong evidence that ensemble perception remains a powerful and flexible heuristic despite the added occlusion. This finding expands work by Haberman and Ulrich (2019), who showed that observers can still extract the average facial expression and identity from a group of faces occluded amodally with black bars.

Experiment 2 also examined feature learning by interleaving trial types together, which discourages participants from learning the features of each stimulus wheel by making that information less predictive or useful. This design allowed us to assess whether participants are maximizing information coming from the available information, a strategy known to be employed by expert face examiners (Carragher et al., 2022), by learning the features of the stimulus set. As illustrated in Figure 3.5, our results indicate that interleaving hurts performance more than adding masks or glasses hurt performance in blocked conditions. In other words, whatever is making ensemble perception robust to the added occlusions is strongly impacted by the interleaved design (i.e., ensemble perception is least resistant to interleaving). This suggests that participants are learning information about the features in the blocked conditions and using those expectations to help form the ensemble statistic. This has implications for how we understand and interpret results from studies using similar face wheels, and would be an interesting area to explore more in future research.

While there has recently been a significant expansion of work looking at the impact of face masks on our ability to process individual faces, very little work has looked at how naturalistic occlusions impact ensemble perception, despite its known role in supporting our visual experience. Future studies, for instance, could incorporate neuroimaging in order to assess whether and how holistic processing is disrupted by the addition of naturalistic occlusions during ensemble perception, as the fusiform face area (FFA) is paramount in holistic face processing while the occipital face area (OFA) plays a greater role in processing facial features rather than whole faces (Tsantani et al., 2021; Liu, Harris, & Kanwisher, 2010).

The results of the current study provide evidence that our visual system is able to extract the ensemble regardless of whether an added occlusion covers the top half or the bottom half of the face. However, further research needs to be done to determine the relative importance of individual facial features in the successful creation of an ensemble percept for faces. Such research would expand a significant literature investigating which facial features are most important for recognition or communication of individual faces (Smith, Cottrell, Gosselin, & Schyns, 2005; McKelvie, 1976; Daview, Ellis, & Shepherd, 1977). It is also unclear whether naturalistic occlusions – which often are objects and have associated meaning themselves – impact ensemble perception for faces in a different way than artificial occlusions like black bars. Future studies could examine the impact of adding naturalistic objects that do not obscure facial features such as hats, tattoos, earrings, etc. on ensemble performance, or strategies observers might use if only half of the ensemble had an added occlusion.

Given its ubiquitousness across many categories of visual information, ensemble perception is widely thought to serve as an important mechanism underlying visual perception and considered a key process used to bypass limited processing bottlenecks. The current study adds to our understanding of ensemble perception by showing its robustness to naturalistic occlusions and highlighting feature learning as a potential mechanism behind extracting the ensemble of faces with limited information. Our findings expand research showing that observers can learn to maximize awareness and use of available features to overcome the impairment face masks cause on single face recognition (Carragher et al., 2022), and contributes to a growing research field examining how ensemble perception for faces is impacted in incomplete or noisy environments (Haberman & Ulrich, 2019).

Chapter 3 is co-authored with Natalia Pallis-Hassani and Timothy F. Brady. The dissertation author was the primary author of this chapter.

References

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. Trends in Cognitive Sciences, 15, 122–131.

Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. Psychological Science, 15, 106–111. doi:10.1111/j.0963- 7214.2004.01502006.x

Alvarez, G. A., & Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. Psychological Science, 19, 392–398.

Alvarez, G. A., & Oliva, A. (2009). Spatial ensemble statistics are efficient codes that can be represented with reduced attention. Proceedings of the National Academy of Sciences United States of America, 106, 7345–7350.

Cabeza, R., Bruce, V., Kato, T., & Oda, M. (1999). The prototype effect in face recognition: Extension and limits. Memory & Cognition, 27(1), 139-151.

Carlaw, B. N., Huebert, A. M., McNeely-White, K. L., et al. (2022). Detecting a familiar person behind the surgical mask: Recognition without identification among masked versus sunglasses-covered faces. *Cognitive Research, 7,* 90. https://doi/10.1186/s41235-022-00440-3

Carragher, D. J., & Hancock, P. J. B. (2020). Surgical face masks impair human face matching performance for familiar and unfamiliar faces. *Cognitive Research: Principles and Implications*, *5*(1), 59. https://doi.org/10.1186/s41235-020-00258-x

Carragher, D. J., Towler, A., Mileva, V. R., White, D., & Hancock, P. J. B. (2022). Masked face identification is improved by diagnostic feature training. *Cognitive Research: Principles and Implications*, *7*(1), 30. https://doi.org/10.1186/s41235-022-00381-x

Cohen, M. A., Dennett, D. C., & Kanwisher, N. (2016). What is the bandwidth of perceptual experience? Trends in Cognitive Sciences, 20, 324–335.

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. Vision Research, 43, 393–404.

Davies, G., Ellis, H., & Shepherd, J. (1977). Cue Saliency in Faces as Assessed by the 'Photofit' Technique. *Perception*, *6*(3), 263–269. https://doi.org/10.1068/p060263

Fischer, J., & Whitney, D. (2011). Object-level visual information gets through the bottleneck of crowding. Journal of Neurophysiology, 106, 1389–1398.

Freud, E., Stajduhar, A., Rosenbaum, R. S., Avidan, G., & Ganel, T. (2020). *The COVID-19 pandemic masks the way people perceive faces* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/zjmr8

Haberman, J.M. & Ulrich, L. (2019). Precise Ensemble Face Representation Given Incomplete Visual Input. *i-Perception 10(1)*. 1-15. https://doi.org/10.1177/2041669518819014

Haberman, J., Lee, P., Whitney, D. (2015) Mixed emotions: Sensitivity to facial variance in a crowd of faces. Journal of Vision 15: 16–16. https://doi.org/10.1167/15.4.16

Haberman, J., Brady, T. F., & Alvarez, G. A. (2015). Individual differences in ensemble perception reveal multiple, independent levels of ensemble representation. Journal of Experimental Psychology: General, 144, 432.

Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. Current Biology, 17, R751–R753.

Haberman, J., & Whitney, D. (2009). Seeing the mean: Ensemble coding for sets of faces. Journal of Experimental Psychology: Human Perception and Performance, 35, 718–734.

Han L, Yamanashi Leib A, Chen Z, Whitney D. Holistic ensemble perception. Atten Percept Psychophys. 2021 Apr;83(3):998-1013. doi: 10.3758/s13414-020-02173-1. Epub 2020 Nov 25. PMID: 33241531.

Haxby, J. V., Ungerleider, L. G., Clark, V. P., Schouten, J. L., Hoffman, E. A., & Martin, A. (1999). The effect of face inversion on activity in human neural systems for face and object perception. *Neuron*, *22*(1), 189–199. https://doi.org/10.1016/s0896-6273(00)80690-x

James W. Tanaka & Martha J. Farah (1993) Parts and wholes in face recognition, The Quarterly Journal of Experimental Psychology, 46:2, 225-245, DOI:10.1080/14640749308401045

Lander, K., & Saunders, G. H. (2023). Face coverings: Considering the implications for face perception and speech communication. *Cognitive Research: Principles and Implications*, *8*(1), 24, s41235-023-00479-w. https://doi.org/10.1186/s41235-023-00479-w

Leib, A. Y., Fischer, J., Liu, Y., Qiu, S., Robertson, L., & Whitney, D. (2014). Ensemble crowd perception: A viewpoint-invariant mechanism to represent average crowd identity. *Journal of Vision, 14,* 26.

Liu, J., Harris, A., & Kanwisher, N. (2010). Perception of face parts and face configurations: an FMRI study. *Journal of cognitive neuroscience*, *22*(1), 203–211. https://doi.org/10.1162/jocn.2009.21203

Maule, J., & Franklin, A. (2015). Effects of ensemble complexity and perceptual similarity on rapid averaging of hue. Journal of Vision, 15(4):6, 1-18.

McKelvie, S. J. (1976). The Role of Eyes and Mouth in the Memory of a Face. *The American Journal of Psychology*, *89*(2), 311. https://doi.org/10.2307/1421414

Sweeny, T. D., Grabowecky, M., Paller, K., Suzuki, S. (2009) Within-hemifield perceptual averaging of facial expressions predicted by neural averaging. *Journal of Vision 9*. 1–11. https://doi.org/10.1167/9.3.2

Smith, M. L., Cottrell, G. W., Gosselin, F., & Schyns, P. G. (2005). Transmitting and decoding facial expressions. Psychological Science, 16, 184–189.

Tsantani, M., Kriegeskorte, N., Storrs, K., Williams, A. L., McGettigan, C., & Garrido, L. (2021). FFA and OFA Encode Distinct Types of Face Identity Information. *The Journal of Neuroscience*, *41*(9), 1952–1969. https://doi.org/10.1523/JNEUROSCI.1449-20.2020

Wolfe, J. M., & Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? Nature Review Neuroscience, 5, 495–501. doi:10.1038/nrn1411

Wolfe, B. A., Kosovicheva, A. A., Leib, A. Y., Wood, K., & Whitney, D. (2015). Foveal input is not required for perception of crowd facial expression. Journal of Vision, 15, 11–11.

Wong, H. K., & Estudillo, A. J. (2022). Face masks affect emotion categorization, age estimation, recognition, and gender classification from faces. *Cognitive Research, 7,* 91. https://doi.org/10.1186/s41235-022-00438-x

Conclusion

In this dissertation, I investigate the impact of prior knowledge on cognitive processes such as memory and perception. The results contribute to the literatures of medical image perception, ensemble perception, and visual working and long-term memory, and open several avenues for future directions in both theoretical and applied domains.

Chapter 1 examined memory performance by non-expert novices and expert radiologists for normal versus abnormal mammography images as a case study in understanding the role of schemas, distinctiveness, and expertise in memory (Schill, Wolfe, & Brady, 2021). We relied on ROC analysis, designed to properly measure memory independent of differences in response criteria and to consider both enhanced memory for seen items as well as the possibility of false alarms. We found that while experts might have access to perceptual encoding benefits, distinctiveness and/or schemas/chunking to enable them to outperform novices, our finding of an extra benefit of expertise for abnormal images is most consistent with a special role of distinctiveness. In other words, the additional knowledge that an expert has compared to a novice gives the expert a boost in memory for abnormal items because they have additional features (i.e., an abnormality), which makes them more distinct in radiologists' memory compared to normal images.

The findings in Chapter 1 open several interesting avenues for future direction. First, future research should expand on the crowd-within effect found in expert populations. For instance, does the crowd-within benefit persist when experts are given unlimited time to process the images? It is also worth noting that long-term memory in this study was tested via a 30-back condition; however, radiologists may be asked to read the same case again up to several months or years later. Thus, it would be interested to see whether the memory advantages for abnormal mammograms found in this study hole true over longer durations or even whether they can be made larger with an even longer delay between the first and second presentation of an image (as found by Vul & Pashler, 2008). Relatedly, continuing to investigate how expertise within a perceptual domain such as radiology develops over time will lead to advancements in both applied and basic literatures of memory, as well as new methods to enhance performance in real-world tasks (Brady et al., 2019; Kundel & La Follette, 1972).

Chapter 2 investigated whether expert's prior knowledge influences their visual perception of images within their domain of expertise (Schill, Gray, & Brady, 2023). We found evidence that radiologists' added knowledge about abnormalities in certain images led to enhanced perception of those images compared to images that did not have the same associated knowledge about an abnormality. Our findings support previous research that has shown that hindsight bias is not only a cognitive, decision-making bias but also one that affects perception (Chen et al., 2020), and is the first to show evidence of a perceptual cognitive bias in expert radiologists reading mammograms.

The presence of a visual hindsight bias in expert radiologists has implications for the legal system and radiologists sued for negligence. Mammography is one of the most common areas within radiology to be sued for negligence (Baker, 2014), and yet there is relatively limited research the effects of visual hindsight bias in radiologists who specialize in mammography. Future research should analyze whether the extent of this bias differs in radiologists depending on the lesion type (e.g., masses, calcifications, and architectural distortions), in addition to difficulty level. Masses and calcifications have very different visual properties (i.e., they vary in size, shape, contrast etc.), which might alter their respective influence on a perceptual bias. It is also unclear whether hindsight bias, whether cognitive or visual, is greater in experts than non-experts within their domain of expertise. To speak towards this ambiguity in the literature, future studies should assess how hindsight bias develops as novices gain experience in their field of expertise. Finally, given the importance of this bias to many applied fields, future research could expand the current literature on ways to reduce or eliminate this bias with a focus on expert populations. Addressing the malpractice lawsuits specifically, future research could contribute to an emerging field that looks at hindsight bias mitigation strategies for juries (Conklin, 2021).

Chapter 3 examined how ensemble perception of faces is impacted when some of the information is obstructed by naturalistic occlusions. Face perception is a useful case study to examine prior knowledge and learning, and face processing is largely considered an area where everyone is a perceptual expert (Tarr & Gauthier, 2000). For instance, we process faces holistically, in a similar way domain expert like radiologists comes to process medical images holistically as they build expertise (Kundel, Nodine, Conant, Weinstein, 2007). Specifically, we asked whether the ability to extract the average facial identity or expression was impacted when the faces had naturalistic occlusions – either face masks or sunglasses, and what strategies participants might employ to combat the added occlusion. The results suggest that while there is some cost, ensemble perception is robust to added occlusions and we can still relatively efficiently extract the mean of a group of faces either varying in identity or expression despite some information being occluded. We also found that that participants appear to be using a strategy that maximizes evidence from the available features to extract the average. Overall, our results appear to be consistent with recent research on how face masks impact perception of individual faces (Carragher et al., 2022) and serve to expand our understanding ensemble perception in circumstances where information is partially obstructed.

The results in Chapter 3 contribute to an expanding literature on how naturalistic occlusions impact face processing and open several interesting avenues for future direction. While the current study could not speak towards whether facial features are weighted differently in their use in creating the ensemble percept, future studies could determine the relative importance of individual facial features in the successful creation of the ensemble representation of faces. It is also unclear how naturalistic occlusions themselves – which often are objects and have associated meaning – impact ensemble perception for faces. For instance, future studies could examine the impact of adding naturalistic objects that do not obscure facial features such as hats, tattoos, earrings, etc. Related to the applied realm, there are several disorders that result in a disruption of holistic face processing. Research that can intentionally disrupt holistic face processing may have the opportunity to investigate why that occurs and strategies individuals can use to combat the inability to process faces holistically.

Medical imaging and face perception are both strong indices of perceptual expertise. Each of these domains has interesting real-world applications, and asking a question from more than one perspective allows more generalizable results to emerge. Research in cognitive psychology has an integral part to play in understanding how individuals and populations make decisions that range from making sustainable choices, diagnosing cancer, to teaching in the classroom. By applying the rigor and control of experimental psychology's methods within more realistic contexts, we can begin to not only speak towards real-world interactions, decisions, and unsolved problems, but also open new avenues to expand our current understanding of cognition.

To conclude, the research program outlined in this dissertation took a deep dive into how experts learn and use learned information to process and perceive complex images. The projects in this dissertation also highlight the applicability of theoretical questions about the nature of perception and memory to our everyday experiences and decisions.

References

Baker, S. R. (2014). U.S. medical malpractice: Some data-driven facts. *Springer International Publishing: Notes of a Radiology Watcher*, 177–179.

Brady, T. F., Störmer, V. S., Shafer-Skelton, A., Williams, J. R., Chapman, A. F., & Schill, H. M. (2019). Scaling up visual attention and visual working memory to the real world. In *Psychology of Learning and Motivation* (Vol. 70, pp. 29–69). Elsevier. https://doi.org/10.1016/bs.plm.2019.03.001

Carragher, D. J., Towler, A., Mileva, V. R., White, D., & Hancock, P. J. B. (2022). Masked face identification is improved by diagnostic feature training. *Cognitive Research: Principles and Implications*, *7*(1), 30. https://doi.org/10.1186/s41235-022-00381-x

Chen, J., Littlefair, S., Bourne, R., & Reed, W. M. (2020). The Effect of Visual Hindsight Bias on Radiologist Perception. *Academic Radiology*, *27*(7), 977–984. https://doi.org/10.1016/j.acra.2019.09.032

Conklin, M. (2021). I Knew It All Along: The Promising Effectiveness of a Pre-Jury Instruction at Mitigating Hindsight Bias. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3962419

Kundel, H. L., & La Follette, P. S. (1972). Visual Search Patterns and Experience with Radiological Images. *Radiology*, *103*(3), 523–528. https://doi.org/10.1148/103.3.523

Kundel, H. L., Nodine, C. F., Conant, E. F., & Weinstein, S. P. (2007). Holistic Component of Image Perception in Mammogram Interpretation: Gaze-tracking Study. *Radiology*, *242*(2), 396–402. https://doi.org/10.1148/radiol.2422051997

Schill, H. M., Gray, S. M., & Brady, T. F. (2023). Visual hindsight bias for abnormal mammograms in radiologists. *Journal of Medical Imaging*, *10*(S1). https://doi.org/10.1117/1.JMI.10.S1.S11910

Schill, H. M., Wolfe, J. M., & Brady, T. F. (2021). Relationships between expertise and distinctiveness: Abnormal medical images lead to enhanced memory performance only in experts. *Memory & Cognition*, *49*(6), 1067–1081. https://doi.org/10.3758/s13421-021-01160-7

Tarr, M. J., & Gauthier, I. (2000). FFA: A flexible fusiform area for subordinate-level visual processing automatized by expertise. *Nature Neuroscience*, *3*(8), 764–769. https://doi.org/10.1038/77666

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, *19*. https://doi.org/10.1111/j.1467-9280.2008.02136.x