Individual Differences in Visual Perceptual Biases

by

Zixuan Wang

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Psychology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor David Whitney, Chair
Professor Richard Ivry
Professor Dennis Levi
Professor Kevin Weiner

Fall 2022

Individual Differences in Visual Perceptual Biases

Abstract

Individual Differences in Visual Perceptual Biases

by

Zixuan Wang

Doctor of Philosophy in Psychology

University of California, Berkeley

Professor David Whitney, Chair

Human perceptual decisions are critically important. They can be socially significant, life-changing or even catastrophic if wrong, such as decisions made by radiologists, Olympic judges, TSA screeners, or vehicle operators. Many of these crucial judgments are made based on the perceptual expertise of individual observers. However, not everyone possesses the same level of expertise, even among professionals. There are two aspects that characterize the individual differences in expertise: variations in sensitivity and variations in bias. Past research has thoroughly investigated differences in sensitivity across many domains. However, the possibility of variations in perceptual bias has been largely ignored.

Here, we filled this gap by investigating the idiosyncratic variations in human visual perceptual biases. In the first study, we identified and characterized the unique perceptual biases in position perception and showed that these biases might originate from variations in visual acuity and they can influence the perceived appearance of objects, suggesting that idiosyncratic perceptual biases could propagate through different levels of visual processing and be associated with each other. To understand whether perceptual biases can have a real-world impact, in the second study, we further tested radiologists with artificial and realistic medical images and found that even expert radiologists were characterized with unique and systematic perceptual biases towards medical images. Given the prevalence of perceptual biases, the third study aimed to discover potential constructive ways to utilize the biases. We showed that when multiple observers make a collective decision, grouping individuals who have uncorrelated and thus independent perceptual biases can improve their combined performance.

Together, these results highlighted the crucial importance of measuring and understanding individual perceptual biases and they have widespread implications for many real-world professions and situations such as driving, radiological or TSA screening and Olympic judges.

Dedication

This dissertation is dedicated to my husband Chengze Fan, my parents Luchuan Wang and Xiaoqin Yang, and especially my grandparents Huizhi Gao and Daosong Wang, for their consistent and unreserved support that inspires me to always be the person I want to be.

Contents

Acknowledgments

Doing a Ph.D. during the historic pandemic is undoubtedly not easy. Looking back, I am sincerely grateful for joining the Whitney Lab and being surrounded and supported by all of the irreplaceable members in the lab over the past four and a half years.

I want to first thank my advisor David Whitney, and three adjunct professors in the lab: Ken Nakayama, Bill Prinzmetal and Erv Hafter. Their brilliant ideas and suggestions have always been a lighthouse to guide me through the ocean of scientific research.

I am also very grateful for my research fellows from the Whitney Lab: Teresa Canas-Bajo, Zhimin Chen, Cristina Ghirardo, Susan Hao, Aki Kondo, Anna Kosovicheva, Allison Yamanashi Leib, Mauro Manassi, Yuki Murai, Jefferson Ortega, Zhihang Ren, Benjamin Wolfe, and Xia Ye. They inspired me with amazing thoughts and provided me with the assistance and encouragement that I would never forget. I also want to thank my collaborator and cheerful cohort Jonathan Tsay for all of the helpful discussions we had. Sincere thanks to my RAs and all of the lab managers with their support. You made this work possible.

Special thanks to the three professors in my dissertation committee members: Rich Ivry, Dennis Levi, and Kevin Weiner for their advice and guidance.

This work was supported in part by the National Institute of Health (grant No. 1R01CA236793-01) to D.W.

Chapter 1

Introduction

Human perceptual decisions are made on a daily basis, and they are crucially important. For example, drivers operate their vehicles on the road majorly based on visual information surrounding them, and expert radiologists perform medical diagnosis mainly based on the perceptual information in the medical images.

Though many of these decisions can be solely dependent on one individual's perceptual expertise, not everyone possesses the same level of perceptual expertise. In the past decade, more and more studies have documented and investigated the variations in the perceptual performance among observers (e.g., Wilmer et al., 2010; R. Wang et al., 2012; Kanai and Rees, 2011; Schütz, 2014; Wexler et al., 2015; Wilmer, 2017; Grzeczkowski et al., 2017; Canas-Bajo and Whitney, 2020; Cretenoud et al., 2020; Cretenoud et al., 2021). These individual differences have been documented from the very lowest level perceptual functions, including contrast sensitivity, motion, and color perception (Schütz, 2014; Kaneko et al., 2018; Emery et al., 2019; Himmelberg et al., 2022) to higher-level object and face recognition skills (Wilmer et al., 2010; Richler et al., 2019; Cretenoud et al., 2020; Cretenoud et al., 2021), and researchers believe that studying these individual differences could be helpful not only for understanding the underlying perceptual mechanism on an individual basis, but also for the real-world improvement on each individual's perceptual performance.

The individual differences in human perceptual expertise can be characterized by two aspects: variations in sensitivity and variations in bias. Past research has thoroughly explored the individual variations in their perceptual sensitivity, suggesting that some individuals are just "better" in one or multiple perceptual functions, ranging from low-level contrast sensitivity to higher-level object or face recognition sensitivity (Peterzell et al., 1995; Wilmer et al., 2010; R. Wang et al., 2012). Studies have also suggested that lower-level sensitivity variations may be originated from the difference in the anatomical structure of primary visual cortex such as cortex thickness and neural population tuning (Kanai and Rees, 2011; C. Song et al., 2015; Himmelberg et al., 2022), while face recognition abilities can be tightly associated with the local white matter properties in the face network such as FFA (Gomez et al., 2015, S. Song et al., 2015).

It may not be surprising that observers can have wildly different perceptual sensitiv-

ity across different tasks, and that is how, usually in real life, experts and amateurs are differentiated. However, is it really just about good or bad?

Recent studies started to demonstrate that human observers are also characterized with idiosyncratic perceptual biases in higher-level object and face recognition tasks (e.g., Afraz et al., 2010; Wexler et al., 2015; Grzeczkowski et al., 2017; Canas-Bajo and Whitney, 2020). They showed that observers can have systematic biases towards objects, faces or visual illusions, and these idiosyncratic variations can swamp the shared biases between individuals. Despite the tendency in showing variations among higher-level object or pattern recognition tasks, our lab's recent work revealed striking idiosyncrasies in a more basic visual function, object localization. We localize objects nearly every moment of every day, making saccades and other eye movements to the text on this page, reaching for a pen or a coffee cup, or appreciating the position of a pedestrian stepping off a curb into the road. Despite the extensive training in localizing objects, individual observers have strong, stable, and consistent idiosyncratic biases in the locations they report objects to be (Kosovicheva and Whitney, 2017).

Despite the prevalence of the idiosyncratic perceptual biases, the relationship between different biases remained unknown. Could these idiosyncratic biases emerge from different levels of visual processing and thus they appear to be irrelevant with each other? Or is it possible that some particular biases could propagate across different levels and thus we could observe an association between different biases? In Chapter 2, we explored this question in the field of spatial biases by investigating the relationship between the individual variations in visual acuity, position perception and the perceived size of objects.

To understand whether these widely-existing idiosyncratic perceptual biases can have any real-world impact, in Chapter 3, we tested expert radiologists' perceptual biases towards medical images and indicated that variations in their perceptual biases could potentially be one of the reasons for explaining the diagnostic performance inhomogeneity across radiologists.

Chapter 4 took a step further by demonstrating that the idiosyncrasies in perceptual biases can be useful to guide our decisions. In this Chapter, we employed three different tasks ranging from low-level visual localization task to higher-level object recognition and emotion tracking tasks, and investigated the possibility that combining the responses from observers who have independent biases could improve their collective decision, which could be of vital importance for many real-world situations such as employing pairs of readers in radiological or TSA screening, as well as more subjective decisions such as performance ratings in Olympic games.

Together, our studies demonstrated the wide existence of idiosyncratic human perceptual biases across a variety of tasks and even among experts, highlighting that measuring and understanding each individual perceptual biases can be a crucial way to improve the accuracies of human perceptual performance.

Chapter 2

Idiosyncratic Perception: A Link Between Acuity, Perceived Position and Apparent Size

Accurately registering the locations of objects is a critical visual function. Most other perceptual functions including pattern and object recognition, as well as visually guided behavior, hinge on first localizing object positions. Position perception is generally assumed to be dictated by retinotopic location, and that may explain a lot of the variance in perceived position. However, perceived position can be biased due to various external factors, such as overt attention, motion and saccadic eye movements (Suzuki and Cavanagh, 1997; Whitney and Cavanagh, 2000; Burr et al., 2001). The impact of these factors can be significant, especially considering the spatial scale at which object recognition and visually guided action happen. A 0.5-degree shift in the location of a pedestrian or car crossing a freeway could result in a catastrophic collision. The scale at which perception and action needs to operate is often very fine, and many factors bias perceived position at a scale that is behaviorally relevant.

In the absence of these external factors, perceived position is often assumed to be uniformly dictated by retinotopic position. However, a recent study challenges this belief and demonstrates that people mislocalize objects idiosyncratically and consistently even without apparent change in the environment (Kosovicheva and Whitney, 2017). The unique biases in object locations were shown to be stable across time when tested after weeks or months, indicating a stable perceptual fingerprint of object location.

Why do people perceive idiosyncratically biased object locations in different parts of the visual field and what are the perceptual consequences of it? Here, we test the possibility that variations in spatial resolution across the visual field might cause the spatial distortions in perceived position. Many researchers have shown that visual acuity varies across the visual field (Low, 1943; Pointer and Hess, 1989; Abrams et al., 2012). Because many models of localization depend implicitly or explicitly on the underlying resolution and homogeneity of spatial coding (Morgan and Regan, 1987; Whitaker and MacVEIGH, 1992; H. Wang and

Levi, 1994;Suzuki and Cavanagh, 1997), it is conceivable that the inhomogeneity in visual acuity could result in an inhomogeneous visual space representation, consisting of areas of contraction (sinks) and expansion (sources). Mislocalization would be one of the natural perceptual consequences of these inhomogeneities.

A further prediction is that if individual observers have inhomogeneous visual acuity and consequential distorted representations of visual space, the biases might be carried along with the visual system so that object representations and appearance may also vary in a predictable and related way. To test this, we also measured whether the perceived size of objects varies at individualized perceptually contracted or expanded regions of visual space.

Experiment 1: Idiosyncratic Visual Space Distortion

Kosovicheva and Whitney (2017) demonstrated that observers have stable and idiosyncratic patterns of mislocalization at different polar angles in the visual field. We hypothesized that this mislocalization pattern reflects distinct distortions of visual space and that it should be observed from the fovea to the periphery (not just at one eccentricity). The purpose of Experiment 1 was to identify whether there are idiosyncratic spatial distortions across the visual field.

Method

Participants

Nine observers (3 females, 2 authors, age range: 19 - 33) participated in this experiment. All subjects were experienced psychophysical participants, and all but the two authors were naïve to the purpose of the study. All subjects reported to have normal or corrected-to-normal vision. Procedures were approved by the Institutional Review Board at the University of California, Berkeley.

Stimuli

Stimuli were presented on a 19-inch gamma-corrected Dell P991 CRT monitor (Dell, Round Rock, TX; 1024 $\times$ 768 pixels resolution, 100 Hz refresh rate). To minimize any off-screen reference (i.e., any visible references outside of the computer monitor including the difference between the monitor frame and the experiment room), the monitor frame was covered by black tape. Visual stimuli were generated using MATLAB (The MathWorks, Natick, MA) and Psychophysics Toolbox (Version 3) (Brainard & Vision, 1997) and the experiment program was run on an Apple Macintosh computer (Apple Inc., Cupertino, CA). Observers viewed the stimuli binocularly at a distance of 40 cm using a chin rest.

We used noise patches as targets for the localization task (See Figure 2.1a). Each noise patch contained random black ($< .001$ cd/m^2, measured by Minolta LS110 Luminance Meter) and white (92.6 cd/m^2) squares (each square was 0.1$\times$0.1 degrees of visual angle [d.v.a.]).

Noise patches were enveloped with a two-dimensional Gaussian contrast aperture (standard deviation: 0.75) and only visible within a circular aperture with a radius of 1.22 d.v.a., Noise patches were shown on one of 5 invisible isoeccentric rings (eccentricity: 2, 4, 6, 8, 10 d.v.a.) and one of 48 angular positions equally distributed with a separation of 7.5 degrees (°) on each ring, which resulted in a total number of 240 possible locations. Angular locations from 0° to 360° correspond to positions starting from the right of the fixation, moving clockwise. The four exactly vertical and horizontal positions at each eccentricity were included.

Procedure

Observers were instructed to maintain fixation throughout the experiment. On each trial, a black ($< .001$ cd/m^2) 0.3-d.v.a. diameter fixation dot was presented at the center of a gray (48.3 cd/m^2) background on the screen. After 1000 milliseconds (ms), a noise patch appeared at a pseudo-randomly chosen location among all 240 locations for 50 ms. Upon the offset of the noise patch, the fixation dot changed to dark gray (30.4 cd/m^2) and 500 ms later, a white (92.6 cd/m^2) 0.45-d.v.a. diameter response dot representing the location of the cursor was superimposed on the fixation and participants freely moved the dot using the mouse to match the location of the noise patch center. The position of the cursor when participants clicked the mouse was recorded as the reported location.

In the experiment, each target location was tested 12 times, so there were 2880 trials in total (12 repetitions $\times$ 48 angular locations $\times$ 5 eccentricities). The whole experiment was separated into 6 sessions (2 repetitions for every location per session, random sequence within session). The time interval between every 2 sessions was on average 1.3 days (standard deviation: 1.6 days).

Data Analysis

For each observer and each session, we first calculated the polar angles of the cursor locations reported by the observer. Then for each target location, the two reported locations from the two trials were averaged within each session. Thus, there were 6 sessions of averaged reported locations. Within each session, the average locations were grouped by 5 eccentricities with each consisting of 48 isoeccentric reported angular locations.

For each session and at each eccentricity, the 48 reported locations were transformed into 48 visual space distortion indices. Since any two physically adjacent target locations at the same eccentricity were separated by 7.5° polar angle, if the angular distance between the two adjacent reported locations in the same session was larger than 7.5°, then the area between them was effectively a region of expanded visual space. On the other hand, if the distance was smaller than 7.5°, the visual space between them was effectively compressed. Thus, when the physical target distance (i.e., 7.5° polar angle) was subtracted from the reported distance (i.e., difference in perceived locations), it yields a visual space distortion index in degrees of polar angle for each location. A more positive index refers to increasing visual space expansion and a more negative index refers to larger visual space compression. Zero

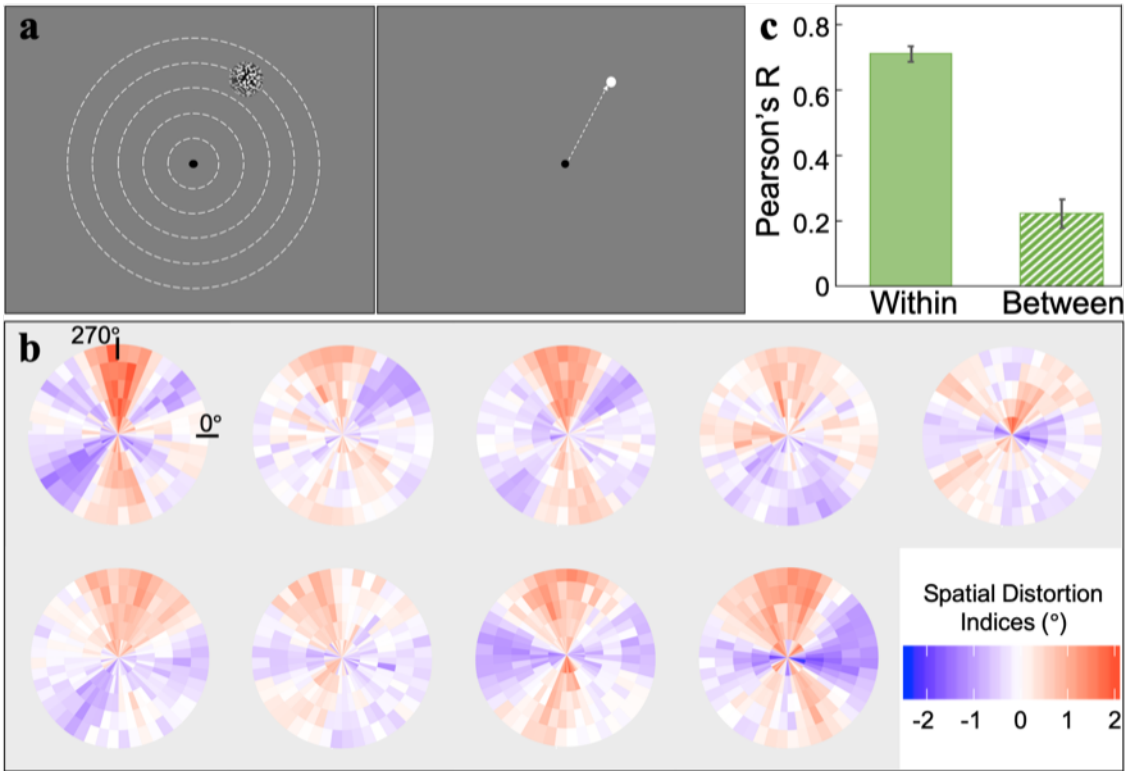

Figure 2.1: Experimental paradigm and results for Experiment 1. (a) Left: Observers fixated at the center and a target was displayed briefly at one of five possible eccentricities (depicted by dashed lines, which were not visible in the experiment). Right: After the target disappeared, observers moved the cursor to match the target's location. (b) All nine individual observers' spatial distortion indices plotted as distortion maps. The color gradient represents the degree of distortion, with blue indicating contracted visual space and red for expanded. See Supplemental Figure S1 for gray-scale luminance-defined distortion maps. (c) Averaged within vs. between-subject correlation calculated by bootstrap procedures (see Experiment 1, Method). There was significantly higher within-observer agreement than between-observer agreement, indicating that each observer had a unique pattern of spatial distortions. The error bars represent the bootstrapped 95% CI.

means no distortion. These resulting 48 distortion indices at each eccentricity within each session were smoothed using a simple moving average method with a window of 45° polar angle. The smoothing is to better characterize a continuous change of distortion across space and to compensate for the discrete spatial sampling given that only 48 locations were tested at each eccentricity. The smoothed distortion indices were averaged across the six sessions at each location (shown as individual distortion map in Fig. 2.1b).

To quantify the idiosyncrasies of the distortion indices, we compare the within-observer to between-observer consistency using a nonparametric bootstrap method (Efron & Tibshirani, 1994). The within-observer consistency is defined as the similarity of the distortion indices across sessions. On each iteration, for every observer, 3 random sessions were sampled without replacement from all of the 6 sessions and the spatial distortion indices for each location were averaged across the 3 sessions as one bootstrapped half. The 3 remaining unsampled sessions were averaged and formed the second half. We z-scored the distortion indices at each eccentricity within each half in order to remove the effect of eccentricity and then correlated the two halves. The Pearson's r value for each observer was transformed into Fisher z value before averaging across observers and the averaged Fisher z was transformed back into Pearson's r value (Fisher transformation, used for all procedures that required averaging correlation values). This procedure was repeated 1,000 times to estimate a 95% bootstrapped confidence interval (CI) for within-observer consistency. Between-observer consistency was estimated similarly. On each iteration, one of the two halves from one observer was correlated with one half from another observer. All possible pairwise between-observer correlations were averaged together. This procedure was also repeated 1,000 times to estimate a 95% bootstrapped CI for between-observer consistency.

To evaluate the significance of the within-observer and between-observer correlations, we also generated permuted null distributions that showed the expected chance correlations from two uncorrelated and permutated halves of the distortion indices. On each iteration, for every observer, all distortion indices were randomly split into two halves (as described previously). One of the two halves was rotated by a random number of positions at each eccentricity to shift the distortion indices away from its original physical positions while simultaneously preserving the spatial relationships between adjacent isoeccentric target locations. The rotated half was then correlated with the other unchanged half from the same observer to estimate a within-observer null correlation. Then, the null correlations were averaged together. This procedure was repeated 10,000 times to estimate a within-observer permuted null distribution. For the between-observer null distribution, on each iteration, the rotated half was correlated with an unchanged half from another observer and all pairwise null correlations between observers were averaged together. This procedure was also repeated 10,000 times. The mean within-observer and between-observer empirical correlations obtained in the bootstrap analyses described above were compared to these null distributions.

To quantify the unique contributions of the distortion indices within each individual rather than a common distortion pattern between observers, we fitted linear regression models and compared the variance explained by different models using a bootstrap test in R (Team, 2013). On each iteration, we first fitted a full model which was formalized as:

$$DI \sim \beta_0 + \beta_1 * self + \beta_2 * others$$

The dependent variable DI (distortion indices) for every target location was calculated by randomly sampling half of the 6 sessions without replacement and averaging the distortion indices across the 3 sampled sessions. Then the 3 remaining unsampled sessions were also averaged to represent the observer-specific spatial distortion pattern (self). The other predictor (others) was the distortion indices averaged across the remaining observers, using only half of the sessions from every observer to minimize the signal-to-noise differences between the two predictors. To estimate how much variance of each observer's distortion indices can be explained by their own distortion pattern (self) versus the other observers' averaged distortion pattern (others), we also fitted two other models, the within-observer and the other-observer models. The within-observer model was formalized as:

$$DI \sim \beta_0 + \beta_1 * self,$$

and the other-observer model was expressed as:

$$DI \sim \beta_0 + \beta_1 * others$$

Unique variance explained by self (i.e., the distortion indices within each observer) was estimated by subtracting variance explained by the other-observer model from that explained by the full model. Variance explained by others (i.e., the averaged distortion indices across other observers) was estimated by subtracting variance explained by the within-observer model from the full model. We also estimated shared variance between self and others by subtracting the unique variances of each of them from the explained variance of the full model. Note that since the shared variance is the common contribution between observers themselves and others, it essentially represents a between-observer similarity in the spatial distortion indices. We repeated this procedure 1,000 times to compare the unique variance explained by within-observer or other-observer distortion indices, or the shared variance between observers.

Results

The bootstrapped within and between-observer similarity is shown in Figure 2.1c. The average bootstrapped within-observer correlation ($r = .71$) was significantly higher than the null correlation expected by chance ($p < .001$, permutation test). This reveals consistent spatial distortion patterns within individual observers. The mean bootstrapped between-observer correlation ($r = .22$) was significantly higher than chance ($p < .001$, permutation test). We further found that within-observer similarity is significantly higher than between-observer similarity ($p < .001$, bootstrap test), suggesting that each individual observer has their own unique spatial distortions that are consistent within themselves and distinguished from other observers (Fig. 2.1c), consistent with a previous study (Kosovicheva & Whitney, 2017). To quantify the unique contributions of within-observer versus between-observer effects, we also fitted linear regression models and compared the unique variance explained

by distortion indices within each observer versus averaged distortion indices across other observers, as well as the shared variance, using a bootstrap procedure (see Data Analysis). Results showed that distortion indices within observers ("self") on average uniquely explained 76.76% of the variance in the full model. Put simply, this means that a particular observer predicts their own pattern of distortions very well. The unique variance that cannot be explained by the observer themselves, and can only be explained by the distortion indices from other observers ("others") was less than 0.1%. This is not surprising: it simply means that other observers' judgments do not have any explanatory power beyond what is already explained by one's own pattern of distortion (i.e., it should be zero). The shared variance between self and others on average explained 23.16% of the variance in the full model. This shared variance is akin to the between-subject similarity in Fig. 2.1c. The unique variance explained by distortion indices within each observer was significantly larger than both the averaged distortion indices across other observers ($p < .001$, bootstrap test, Bonferroni corrected $\alpha_B = .025$) and the shared variance between these two predictors ($p < .001$, bootstrap test, $\alpha_B = .025$). Since the shared variance between self and others captures the shared contributions from every observer versus all other observers, this is essentially a between-observer effect. The regression models show that idiosyncratic observer-specific biases are the major contributor to the spatial distortions, rather than a common spatial bias among observers.

Experiment 2: Associate Individual Distortion Fingerprints with Visual Acuity

In Experiment 1, we demonstrated that individual observers have unique visual space distortions. Where do these idiosyncratic spatial biases emerge? Given that previous studies have revealed substantial heterogeneity in visual acuity across the visual field (Low, 1943; Pointer and Hess, 1989; Abrams et al., 2012), it is plausible that the spatial distortions emerge as a consequence of this. Therefore, in Experiment 2, we measured Vernier acuity (Levi & Klein, 1985) at different spatial locations to assess the potential association between spatial distortions and variations in visual resolution. We used Vernier acuity task because unlike other acuity measurements such as grating acuity, Vernier acuity, also called hyperacuity (Westheimer, 1975), exceeds the limits imposed by the maximal cone density on the retina, and has been shown to measure acuity at a cortical level (Duncan & Boynton, 2003).

Method

Participants

Seven observers (2 females, 2 authors, age range: 19 - 33) who had participated in Experiment 1 participated in the second experiment. All subjects were experienced psychophysical observers, and all but the two authors were naïve to the purpose of the study. All subjects

reported to have normal or corrected-to-normal vision. Procedures were approved by the Institutional Review Board at University of California, Berkeley.

Stimuli

Stimuli were generated by the same software and displayed on the same monitor as Experiment 1. Observers were tested binocularly with a viewing distance of 40 cm fixed by a chin rest. Two long (1.5 d.v.a.) and thin (50 arcsec) white lines (92.6 cd/m^2) oriented towards the center of the screen were shown on a gray background (48.3 cd/m^2). On half of the trials, the outer line was positioned clockwise from the inner line and counter-clockwise for the other half of the trials (Figure 2.2a shows the counter-clockwise condition). Five possible spatial misalignments between the two lines were deployed; the range of misalignments was customized for each observer after a practice block at the beginning of the experiment. Among all observers, the spatial misalignments were between 0.3 to 10 arcmin of visual angle. As a result, there were 10 possible Vernier stimuli (2 possible positional relationships $\times$ 5 possible spatial misalignments). The center of the range of Vernier stimuli was located at one of 8 angular positions (45° spacing, starting with 20° polar angle), on an invisible isoeccentric ring with a radius of 6 d.v.a..

Procedure

On each trial, a black ($<$ .001 cd/m^2) 0.3-d.v.a. diameter fixation dot was presented continuously at the center of a gray screen to help observers maintain fixation during every trial. After 1,000 ms, two Vernier lines were shown for 500 ms on a pseudo-random location selected from the 8 possible positions as described above. The 10 different Vernier stimuli were presented in a pseudo-random sequence, such that every appearance was guaranteed to be displayed the same number of trials on every location. After the Vernier stimuli disappeared, observers responded by either pressing left key or right key to indicate the spatial relationship of the two lines (yes/no design). If the more eccentric line appeared to be relatively more counter-clockwise (clockwise), observers were instructed to press the left (right) arrow key. Observers were instructed to respond as accurately as possible and they had unlimited time. They were also told to fixate on the fixation dot at all times. In the experiment, each of the 10 Vernier misalignments was repeated 20 times, resulting in 200 trials at every tested location. Since we presented stimuli at 8 possible positions, there were 1,600 trials in total. The whole experiment was separated into 16 blocks and each block contained 100 trials. Observers were encouraged to take a break after each block.

Data Analysis

For every Vernier misalignment at a given location, we calculated the proportions in which observers reported that the outer line was shifted more clockwise than the inner line. Then we fitted the proportion of clockwise responses with a logistic function using a least-squares procedure. The just-noticeable difference (JND) was estimated by taking half of the distance

between the Vernier misalignments that gave 25% and 75% clockwise responses on the best-fit logistic function. This fitting procedure and JND estimation were conducted separately for each location; thus, we obtained 8 JND values from each observer (See Figure 2.2b for a representative subject).

To estimate the spatial distortion indices at these 8 locations, for each observer, each of the eight locations tested in this experiment was rounded to the two nearest locations that had their own corresponding distortion indices in Experiment 1 at the eccentricity of 6 d.v.a. and the two distortion indices were averaged as a proxy of the spatial distortion at each of the eight locations tested in Experiment 2. We then calculated the Pearson's correlation between spatial distortions indices (Experiment 1) and Vernier acuity JNDs (Experiment 2) for every observer separately. For each of the 7 observers, 8 Vernier acuity JND values were correlated with the corresponding 8 spatial distortion indices estimated from Experiment 1. This yielded in total 7 Pearson's r values, which were Fisher transformed and averaged to estimate the mean correlation. We also performed a bootstrap procedure on these 7 correlation values to test whether mean correlation value was biased by extreme observer(s). On each iteration, we randomly sampled 7 correlation values with replacement from the 7 empirical Pearson's rs and averaged the 7 Fisher-transformed correlations. We repeated this procedure for 1,000 times and estimated the 95% bootstrapped CI for the mean correlation among observers.

To examine whether individual differences play a role in the relationship between Vernier acuity JNDs and spatial distortions, we fitted a linear mixed-effect regression on the data. The model could be expressed as:

$$JND_{i,j} = \beta_0 + \beta_1 * DI_{i,j} + \gamma_i$$

$JND_{i,j}$ was the Vernier acuity JND calculated for the j^{th} location of the i^th observer. $DI_{i,j}$ was the distortion index calculated for the j^th location of the i^{th} observer. γ_i represented the random effect for the i^{th} observer. We also constructed a simple linear model without accounting for the random effect of observers, which could be formalized as below.

$$JND_{i,j} = \beta_0 + \beta_1 * DI_{i,j}$$

A likelihood ratio test was performed between the two models to compare the goodness-of-fit of the models with or without the random effect of observers.

Results

The change of Vernier acuity as a function of different angular locations for each individual observer is shown in Supplemental Figure S2. Figure 2.2c visualizes the relationship between Vernier acuity and spatial distortion with individual differences removed (i.e., Vernier acuity JNDs and spatial distortion indices were z-scored within each observer and then plotted into the same figure as a "super subject"). We are aware that any analysis based on the "super subject" data may be subject to the problem of pseudo-replication (Lazic, 2010),

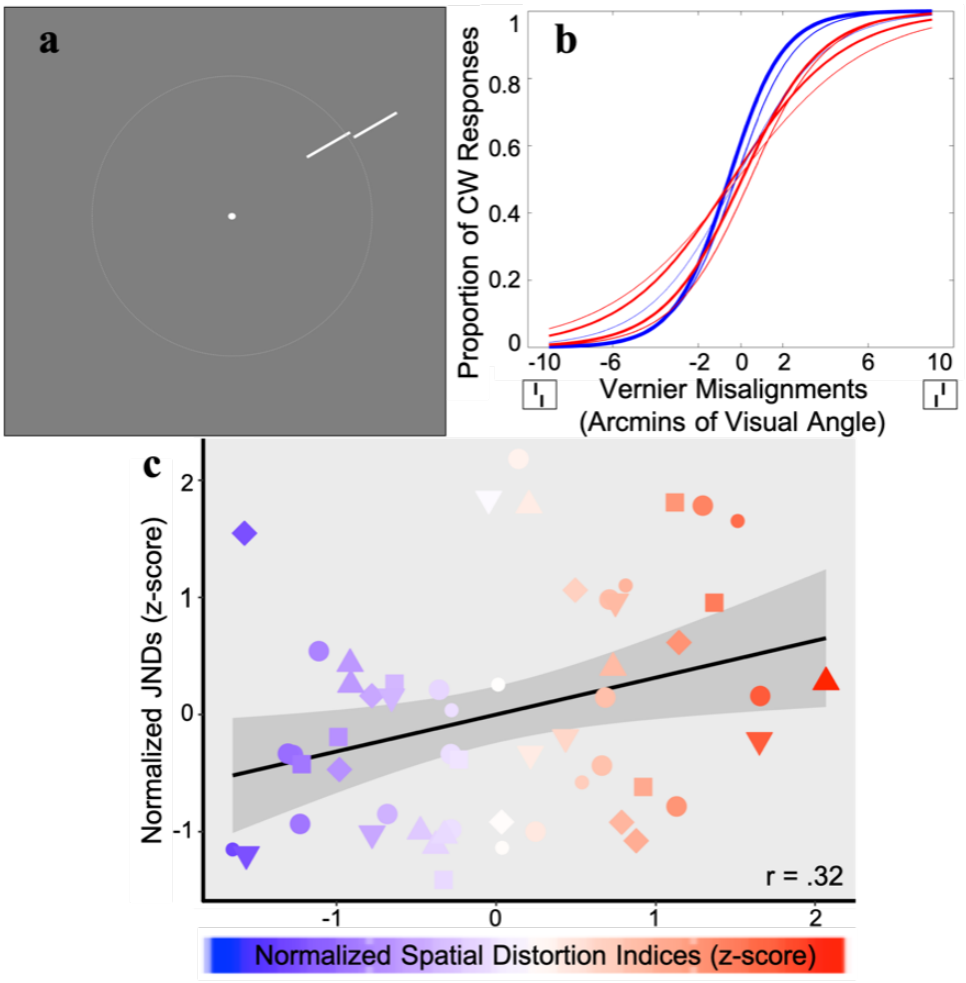

Figure 2.2: Paradigm and results for Experiment 2. (a) Example Vernier stimuli. Stimuli were shown on an invisible circle with a radius of 6 d.v.a. (dotted line). (b) Psychometric curves fitted for one representative subject. The colors of the lines correspond to locations in the visual field that had different types of distortion (blue for compression and red for expansion, as measured in Experiment 1); the weights of the lines correspond to the intensity of the distortion. (c) A visualization of the correlation between Vernier acuity JNDs (Experiment 2) and corresponding spatial distortion indices (Experiment 1) for all 7 observers (represented by different shapes and sizes) collapsed into a single super-subject. Blue shading indicates more contracted locations and red shading indicates more expanded locations from Experiment 1. The black line is the best-fitted linear regression line based on this super-subject data and the gray shaded area represents the 95% CI of the linear fit. There was a significant positive correlation between the degree of visual space distortion and Vernier JNDs, indicating that better acuity was found at perceptually contracted visual space compared to expanded visual space.

which is why we constructed a linear mixed effect regression (see Data Analysis section) and calculated individual correlations for each subject (Fig. S3).

Across the individual subjects, there was a significant correlation between the JNDs calculated in Experiment 2 and the corresponding spatial distortion indices at the same locations taken from Experiment 1 (mean Pearson's $r = 0.34$, 95% bootstrapped CI: [0.06, 0.56]; Fig. S3). This indicates that the spatial distortion index increased with increasing JND: the locations in the visual field where acuity was better tended to be perceived as more compressed, compared to the regions of spatial expansion where acuity was coarser. Individual observer correlations are shown in the supplementary materials (Fig. S3).

The results of the linear mixed-effect model also confirmed that there was a significant and positive relationship between Vernier acuity and spatial distortions, with a fixed effect coefficient of 0.63 (standard error: 0.23, $F(1, 48) = 7.59, p < .01$). This confirms the analysis above, indicating that higher acuity is associated with compression of perceived space. Importantly, compared to the model without the random effect of observers, the mixed-effect model was significantly better (likelihood ratio test, $\chi^2(1) = 39.99, p < .001$). In other words, accounting for individual differences significantly increased the performance of the mixed-effect model, which suggests that the association between Vernier acuity and spatial distortion is characterized by observer specific idiosyncrasies.

Experiment 3: Distorted Visual Space Modifies Object Appearance

In the previous experiments, we found unique spatial distortions for individual observers (Experiment 1) and also discovered a potential source of the biases based on variations in visual acuity (Experiment 2). While acuity might be determined by early visual cortical processes (Duncan & Boynton, 2003)), position perception, arguably, emerges in extrastriate visual areas (McGraw et al., 2004; Fischer et al., 2011; Maus et al., 2013). This hints at the possibility that idiosyncratic biases in perceived position (Experiment 1) might be inherited along the visual hierarchy to other later visual processing stages, such that they actually change the appearance of objects. We tested this question in Experiment 3 by measuring the perceived size of objects presented at different spatial positions.

Method

Participants

3 observers (1 female, 2 authors, age range: 23 - 33) who have participated in Experiment 1 and 2 participated in the current experiment. Experiment 1 and 2 have demonstrated that the association between acuity and spatial distortions is observer-specific, so a dense spatial sampling within a single subject will be necessary and sufficient to establish the relationship between variations in perceived size and spatial distortions. Therefore, in Experiment 3,

we recruited fewer participants but with more locations tested for each participant. One observer was naïve to the purpose of the study. All subjects reported to have normal or corrected-to-normal vision. Procedures were approved by the Institutional Review Board at University of California, Berkeley.

Stimuli

Stimuli were generated with the same software and displayed on the same monitor as Experiment 1 and 2. Observers were tested binocularly with a viewing distance of 40 cm fixed by a chin rest. Stimuli consisted of "arcs" drawn from an invisible circle with a radius of 6 d.v.a. (Fig. 2.3a). The arc stimuli were shown at one of 20 angular positions equally distributed on the circle. Angular positions were separated by 18°, starting from 9°. There were 6 possible arc measures: 16.82°, 16°, 15.55°, 14.45°, 14°, 13.18°; the mean arc measure was 15°. Arc width was 4 arcmin of visual angle.

Procedure

We used the method of single stimuli (McKee et al., 1986;Morgan et al., 2000) to estimate the perceived size of objects across the visual field. Observers were instructed to maintain fixation throughout each trial on a black ($< .001$ cd/m^2) 0.3-d.v.a. diameter fixation dot at the center of a gray (48.3 cd/m^2) screen. After 1,000 ms, a white (92.6 cd/m^2) arc (see descriptions above), with a pseudo-random length selected from the 6 possible lengths, was presented at one of the 20 predetermined positions. The arc was displayed for 500 ms and then it disappeared from the screen. Upon its offset, observers pressed either the left or right arrow key to indicate whether the presented arc was shorter or longer than the average of all seen arcs, regardless of stimulus location. Observers had unlimited time to make the key-press response and they were told to try to be as accurate as possible.

Observers were exposed to all possible arc lengths during a practice block, which was completed before the start of the experiment. The procedure in the practice was almost identical to the actual experiment, except that there were only 120 trials with each of the 6 arc lengths repeated 20 times. Also, observers received feedback in the practice block to speed their learning of the set mean; their accuracy was around 80% correct at the end of the practice session. In the actual experiment, no feedback was provided.

The whole experiment was divided into 2 sessions. In each session, observers finished a practice block and then an experiment block. In each experiment block, there were 1,200 trials in total, with the condition that each possible length (6 in total) was presented 10 times at each possible location (20 in total). Observers were encouraged to take a rest after each set of 150 trials. The two sessions of 1200 trials each were collected on separate days, for a total of 2400 trials.

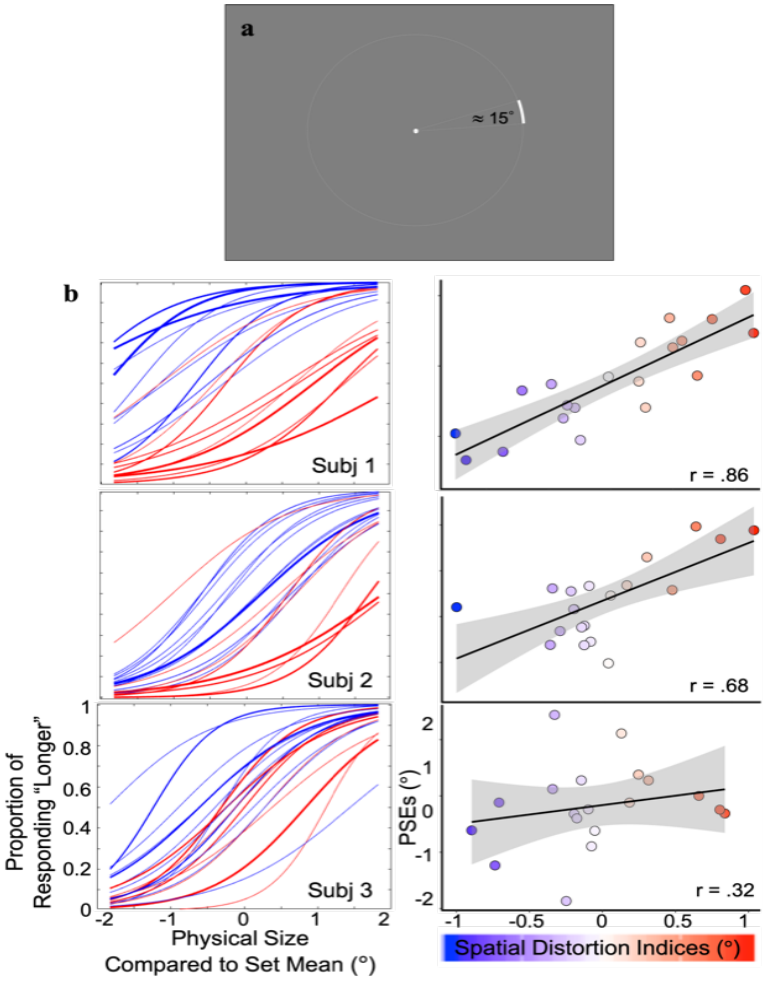

Figure 2.3: Experiment 3 paradigm and results. a) Example arc stimulus used in the experiment. On each trial, an arc was presented at one of the 20 locations separated by 18° at an eccentricity of 6 d.v.a.. Upon the offset of the arc, observers responded whether it was shorter or longer than the average. b) Top: Psychometric curves fitted for all three observers (Subjects 2 and 3 are authors). The abscissa represents the angular size difference between the presented and the mean arc. Colors of the lines correspond to different distortions obtained from individual subjects in Experiment 1 (blue for compression and red for expansion) and the weights of the curves correspond to the intensity of the distortion. Bottom: The association between the idiosyncratic visual space distortions (abscissa, from Experiment 1) and perceived size (ordinate) for each subject. The color scale corresponds to the extent of compression or expansion quantified by the spatial distortion indices in Experiment 1. The positive correlations indicate that in regions of contracted (expanded) visual space, objects were perceived to be larger (smaller) than their actual size. The gray shaded area around the regression line demonstrates the 95% CI of the linear regression fit.

Data Analysis

To estimate the perceived size of the arc at each location, we calculated the proportion of trials in which the observer responded "longer than average" for each length of the arc. Then these proportions were fitted to a logistic function using a least-squares procedure. The point of subjective equality (PSE), which represents the perceived set mean, was defined as the arc length at which the proportion of "longer" responses was 50% on the best-fit logistic function. A larger PSE represents a smaller perceived object size. We repeated this procedure for 20 tested locations; thus, we obtained 20 PSE values for each participant (Figure 2.3). The change of PSE values as a function of different angular locations for each individual observer is shown in the Supplemental Figure S4. To confirm the test-retest reliability, we estimated these 20 PSE values separately for each of the two sessions. The PSEs across sessions were averaged to test the relationship between the spatial distortions (Experiment 1) and the perceived object size (Experiment 3).

For each observer, each of the 20 locations in the current experiment was rounded to the two nearest corresponding locations tested in Experiment 1. The corresponding averaged distortion indices from Experiment 1 were correlated with the PSEs obtained in this experiment. To test whether the correlation yielded for each observer was significantly higher than chance, we performed a permutation test. For each observer, we randomly shuffled the location labels of the PSEs and then correlated the shuffled PSEs with the distortion indices. This procedure was repeated 10,000 times to generate a null distribution and estimate the significance of the empirical correlation for each observer relative to this null distribution.

We analyzed the Pearson's correlation between the PSEs from same locations but from different sessions to estimate the within-observer consistency of the spatial heterogeneity in perceived size for every participant. Between-observer consistency was also estimated by calculating pairwise correlations and averaging across all possible pairs.

Results

We found that across all three observers, spatial distortion indices were closely associated with perceived size (Fig. 2.3), with Pearson's correlations of .86, .68 and .32 respectively for three observers ($p < .001, p < .001$ and $p = .13$ respectively, permutation test; Fisher's combined probability test: $p < .001$). This suggests that object appearance is altered depending on the local distortions in spatial coding. At perceptually compressed regions of the visual field (i.e., smaller distortion indices), objects are perceived to be larger, while the apparent size of objects is smaller at apparently expanded regions of the visual field. The within-observer consistency estimated by correlating PSEs across sessions, within each observer, showed that each observer had a stable (Pearson's correlations: .93, .94, .87, $ps <$ .001, last two are the authors) and observer-specific (compared to a much lower between-observer, $r = 0.39$) spatial heterogeneity in the perceived sizes of objects throughout the visual field. Apparent object size is therefore affected by the unique spatial distortions

across each observer's visual field, suggesting that idiosyncratic spatial biases are inherited along the visual hierarchy.

General Discussion

In the present study, we found that observers are characterized by idiosyncratic spatial distortions, and we showed that these biases may arise from inhomogeneous spatial acuity across the visual field. These biases are then inherited along the visual hierarchy such that they change the appearance of object size. In Experiment 1, we found that different regions of the visual field in different observers are either effectively compressed or expanded. We demonstrated that these spatial distortions for each individual observer cannot be simply attributed to common biases among individuals. To trace the possible source of the distortions, Experiment 2 tested whether Vernier acuity varies at regions of the visual field that are perceptually compressed versus expanded. The results supported our hypothesis that spatial acuity is tightly linked to subject-specific spatial distortions. In particular, better acuity is associated with spatial compression and worse acuity is associated with a perceptual expansion of space. We also found that the association between acuity and the spatial distortion indices is observer-specific, which confirms the idiosyncratic nature of the spatial distortions. In Experiment 3, we asked whether the same idiosyncratic spatial biases can be inherited at higher levels of visual processing, such that they change the appearance of visual objects. We found a stable and observer-specific spatial heterogeneity of perceived size of objects, which can be predicted by the idiosyncratic spatial distortions. Within regions of the visual field that were distinctly distorted (measured from biased position perception from Experiment 1), the apparent size of objects changed predictably: Objects were perceived to be larger at locations with perceptually contracted spatial representation and smaller at expanded locations.

The individual differences in spatial distortions and apparent object size do not contradict known between-subject spatial biases or spatial anisotropies of vision such as the oblique effect (Appelle, 1972). In fact, we did find a significant between-observer correlation in Experiment 1. And, when we averaged the idiosyncratic distortions across observers, we replicated other common spatial biases that have been reported by researchers (Low, 1943; Pointer and Hess, 1989; Abrams et al., 2012). For example, on average, observers tend to perceive visual space as expanded along the vertical meridian and compressed along the horizontal meridian (for average distortion map, see Supplemental Figure S5). Since we found that better acuity underlies compressed visual areas and relatively worse acuity at expanded regions, this group-level result aligns with previous studies that showed a horizontal-vertical anisotropy (Low, 1943). Note that even along the vertical meridian, our results revealed an upper-lower visual field asymmetry, with the vertical axis in the upper visual field being more expanded. This also replicates the vertical meridian asymmetry in both human and non-human primates (Previc, 1990; Abrams et al., 2012). A more recent study carried out by Greenwood and his colleagues 2017 found that crowding zones and saccadic error zones are

on average larger along the vertical than horizontal meridian and that there are variations between observers, which is consistent with what we showed in the current study. Our study goes beyond previous ones by showing that the individual differences in acuity, perceived position, and perceived size are unique to the individual observer, are highly reliable, and are not due to a group level effect. Therefore, we believe that our current study reveals a fundamental idiosyncrasy underlying each observer's visual system and that our findings can potentially explain part of the common spatial biases reported in past research.

Despite the shared, between-subject, distortions and consistency (Fig. S1), it is worth noting that the within-subject effects are sufficiently consistent and strong that they can swamp the between-subject effect. The within-subject correlation is not only higher (Fig. 2.1) but also contributed more unique variance than shared variance when explaining each subject's individual distortion indices (Experiment 1, Results). Experiment 2 showed that the association between distortions and Vernier acuity is significantly observer-specific (Experiment 2, Results). This suggests that individual differences in spatial resolution, as measured by Vernier acuity, are also substantial (Fig. S3). Experiment 3 again revealed an observer-specific spatial heterogeneity of perceived size, which is reliably predicted by the idiosyncratic spatial distortions within each observer (Experiment 3, Results).

The within-subject effects also appear to be fairly stable over time, consistent with prior work (Kosovicheva & Whitney, 2017). The time span between Experiment 1 and 2 was 1 2 months, and the time span between Experiment 1 and 3 was around 11 months. Given the link we found between the biases estimated from different experiments, this suggests some degree of stability in the within-subject idiosyncrasies, which are consistent with the temporally stable spatial distortions that were previously reported (Kosovicheva & Whitney, 2017). Overall, the rich variation between observers is in line with previous studies that showed substantial and stable idiosyncratic biases in the recognition of objects and patterns and motor decisions (Abrams et al., 2012; Schütz, 2014; Wexler et al., 2015; Grzeczkowski et al., 2017).

What causes the idiosyncratic acuity found in Experiment 2? One possibility is that variations in visual acuity arise from spatial inhomogeneities in early visual cortical processing. Previous studies have revealed that there are individual differences in the anatomical structure of human primary visual cortex (Kanai & Rees, 2011), and that these are associated with differences in neural population tuning and perceptual performance in different visual tasks (C. Song et al., 2015; Moutsiana et al., 2016). Therefore, it is plausible that even at the same eccentricity within individual observers, anatomical profiles corresponding to distinct regions of the visual field may be fundamentally heterogeneous and this creates an intrinsic spatial bias from early stages of the cortical visual hierarchy. When visual information is passed to later visual areas, local variations are inherited and neurons in those higher-level visual regions can be impacted through neural undersampling processes (Afraz et al., 2010). Although speculative, our results suggest that this inheritance happens both in the ventral occipitotemporal visual pathway for object vision and the parietal pathway for spatial vision (Mishkin et al., 1983); this includes higher-level regions such as inferior temporal cortex, which has been linked to the perceived size of objects (Cohen et al., 1994), and the posterior

parietal cortex, which is responsible for spatial representation and head-centered localization (Andersen et al., 1997).

How can the fundamental anatomical structure of our primary visual cortex be inhomogeneous? One possibility is random variation in receptive field number, size, and/or density in any isoeccentric location of the visual field. Another possibility is that inhomogeneities are introduced through development. For example, heterogeneous spatial resolutions could be the aftermath of infant astigmatism (Mohindra et al., 1978). Although significant astigmatism found in infancy typically disappears with age (Atkinson et al., 1980), such early astigmatism could introduce or shape inhomogeneities in visual cortical density and/or receptive field properties. It is also plausible that the idiosyncrasy we find may be determined naturally by genes and shaped by neuron migration, cortical connectivity and brain folding in early stages of life. For example, individual differences in the ability to recognize faces were found to be largely genetically driven through twin studies (Wilmer et al., 2010). Future research on the visual development of individuals is needed to understand why humans develop such an idiosyncrasy in vision.

Although stable, consistent, and accurate spatial localization is frequently assumed to be a simple product of retinotopic position, our results challenge this belief and demonstrate that every individual is characterized by idiosyncratically distorted visual space. These distortions may be induced by heterogeneous spatial acuity across the visual field, and can influence visual appearance of object size, which suggests that these idiosyncratic fingerprints may be driven by variations in early visual cortex and are inherited along the visual hierarchy.

In Chapter 2, we discussed the potential origins of the idiosyncratic biases in object localization and also showed these biases can predict the perceived size of objects. However, it remained unknown whether these perceptual biases would exist with real-world, complex stimuli and also within experts. To answer this question, in the next Chapter, we investigated the variations in the perceptual biases towards medical images from expert radiologists.

Chapter 3

Idiosyncratic biases in the perception of medical images

Medical image perception is fundamentally important for decisions that are made on a daily basis by clinicians in fields ranging from radiology and pathology to internal medicine (Samei & Krupinski, 2018). At a fundamental level, the kinds of decisions that are made depend on the perceptual information that is available to these clinicians (Kundel, 2006; Samei and Krupinski, 2009; Krupinski, 2010). This hinges largely on the clinicians' basic perceptual abilities as human observers (Kundel, 1989; Quekel et al., 1999; Donald and Barnard, 2012), as well as their specific training and experience (Fletcher et al., 2010; Theodoropoulos et al., 2010; Sha et al., 2020).

It has been known for decades that radiologists have significant individual differences in their diagnostic performance (Elmore et al., 1994; Feldman et al., 1995; Beam et al., 1996; Elmore et al., 1998; Tan et al., 2006; Elmore et al., 2002; Lazarus et al., 2006; Elmore et al., 2009; Sonn et al., 2019; Pickersgill et al., 2019). For example, radiologists vary in the accuracy of their mammography reading (e.g., Feldman et al., 1995; Beam et al., 1996; Tan et al., 2006; Elmore et al., 2002). Similar results were found in prostate magnetic resonance imaging screening (e.g., Sonn et al., 2019; Pickersgill et al., 2019). Some studies suggested that these strong individual differences are due to variation in radiologists' training (e.g., Linver et al., 1992; Berg et al., 2002; Van Tubergen et al., 2003), as well as their experience level (e.g., Herman and Hessel, 1975; Elmore et al., 1998; Manning et al., 2006; Molins et al., 2008; Rosen et al., 2016). Other studies proposed that some differences may be due to the strategies adopted by radiologists (Kundel and La Follette Jr, 1972; Kundel et al., 1978; Krupinski, 1996). For example, radiologists tend to follow two main search strategies. "Drillers" keep fixation on a certain area, and scroll through depth, whereas "Scanners" scan an entire image before moving to the next one (Drew et al., 2013; Mercan et al., 2018).

In recent years, more and more studies have documented and investigated the individual variations in the perceptual performance among groups of untrained observers (e.g., Wilmer et al., 2010; R. Wang et al., 2012; Kanai and Rees, 2011; Schütz, 2014; Wexler et al., 2015; Wilmer, 2017; Grzeczkowski et al., 2017; Z. Wang et al., 2020; Canas-Bajo and Whitney,

2020; Cretenoud et al., 2020; Cretenoud et al., 2021) and a few studies also investigated the perceptual abilities among clinicians including radiologists (see Waite et al., 2019 for a review; Smoker et al., 1984; Corry, 2011; Birchall, 2015; Langlois et al., 2015; Sunday, Donnelly, & Gauthier, 2017, 2018). Typical human observers actually have substantial individual differences in their perceptual abilities and biases (for reviews, see Wilmer, 2017; Grzeczkowski et al., 2017; Mollon et al., 2017). These individual differences have been documented from the very lowest level perceptual functions, including localization, motion, and color perception (Schütz, 2014; Wexler et al., 2015; Kosovicheva and Whitney, 2017; Kaneko et al., 2018; Emery et al., 2019; Z. Wang et al., 2020) to higher-level object and face recognition skills (Wilmer et al., 2010; Richler et al., 2019; Canas-Bajo and Whitney, 2020; Cretenoud et al., 2020; Cretenoud et al., 2021). For example, we localize objects nearly every moment of every day, making saccades and other eye movements to the text on this page, reaching for a pen or a coffee cup, or appreciating the position of a pedestrian stepping off a curb into the road. Despite the extensive training in localizing objects, individual observers have strong, stable, and consistent idiosyncratic biases in the locations they report objects to be (Kosovicheva and Whitney, 2017; Z. Wang et al., 2020).

Another example of striking individual differences is face recognition, which varies substantially between observers (Duchaine and Nakayama, 2006; Russell et al., 2009; Wilmer et al., 2010; R. Wang et al., 2012; Russell et al., 2012; Bobak et al., 2016). For example, so called "super recognizers" can match the identity of random photographs of children to their corresponding adult photographs, whereas those with prosopagnosia often cannot recognize the identity of faces, even themselves or loved ones (Duchaine and Nakayama, 2006; Klein et al., 2008;Russell et al., 2009). These individual differences arise despite extensive training and everyday experiences observers have with faces, and despite the many brain regions and networks devoted to the processing of faces (Kanwisher et al., 1997; Gauthier et al., 2000; Haxby et al., 2001). Holistic face recognition, inversion effects, fractured faces, and other kinds of illusions demonstrate the richness, sophistication, and specialization that we have for recognizing faces (Moscovitch et al., 1997; Farah et al., 1998; Maurer et al., 2002; Rossion, 2013). Still, despite all of that training and exposure, human observers have wildly different face recognition abilities. A great deal of the individual differences in human visual perception might be explained by genetic variations (Wilmer et al., 2010; Q. Zhu et al., 2010; R. Wang et al., 2012; Z. Zhu et al., 2021), but other individual differences are due to training and experience (Germine et al., 2015; Sutherland et al., 2020; Chua and Gauthier, 2020).

This body of recent perceptual research provides important insights for the idiosyncrasies among radiologists. A first possibility is that there are differences in perceptual sensitivity including visuospatial skills and novel object recognition abilities between clinicians (Smoker et al., 1984; Corry, 2011; Birchall, 2015; Langlois et al., 2015; Sunday, Donnelly, & Gauthier, 2017, 2018), just like individuals vary in their sensitivity when recognizing faces (Wilmer et al., 2010). This has been the major focus of previous studies investigating individual differences in radiologist perception (see Waite et al., 2019 for a review). These differences in sensitivity could be a natural consequence of variability in experience and training (Herman and Hessel, 1975; Linver et al., 1992; Elmore et al., 1998; Berg et al., 2002; Van Tubergen et

al., 2003;Manning et al., 2006; Molins et al., 2008; Rosen et al., 2016). Other potential factors include genetic variations, and are not unexpected and could be superseded by training (Bass & Chiles, 1990). A second non-exclusive possibility is that there are differences in perceptual biases between different clinicians. For example, clinicians might systematically and consistently misperceive textures, colors, shapes and locations in different ways, as it is known to occur in untrained observers (Schütz, 2014; Wexler et al., 2015; Kosovicheva and Whitney, 2017; Kaneko et al., 2018; Emery et al., 2019; Z. Wang et al., 2020; Canas-Bajo and Whitney, 2020; Cretenoud et al., 2020; Cretenoud et al., 2021).

Whether there are idiosyncratic perceptual biases that clinicians bring to medical image recognition tasks has not been closely studied, but any biases that exist could influence accuracy, diagnostic errors, etc., even if perceptual sensitivity was constant. Conversely, the individual differences in perceptual sensitivity among radiologists (Birchall, 2015; Sunday, Donnelly, & Gauthier, 2017, 2018) do not predict that there are necessarily systematic idiosyncratic perceptual biases. In fact, there may be no idiosyncratic biases in perception despite the individual differences in accuracy. This is worth reiterating: individual differences in sensitivity need not be the same as individual differences in bias (even if they could be correlated suggested by Wei and Stocker, 2017). Therefore, the question of idiosyncratic biases in clinician perception remains unknown and untested in prior literature.

One reason we believe that investigating perceptual biases (as opposed to sensitivity) was difficult in prior research is that the stimuli used were natural (clinical settings) and therefore not easily or well controlled. Hence, it is almost impossible to measure systematic perceptual biases in radiologists in those studies. In order to measure these idiosyncratic biases in the medical image perception performance of radiologists, we need controlled stimuli and experiments. The goal of this study was to test for idiosyncratic perceptual biases in a group of radiologists with controlled visual stimuli. We also compared the radiologists' results to a comparable group of naïve participants who were untrained and inexperienced with medical images.

Experiment 1

Raw data for Experiment 1 were obtained from a previously published experiment on perceptual judgments by radiologists and untrained non-clinical observers (Manassi et al., 2021).

Methods

Participants

Fifteen radiologists (4 female, 11 male, age: 27-72 years) and eleven untrained college students (7 female, 4 male, age: 19-21 years) were tested in the experiment. Radiologists participated on site at RSNA annual meeting and college students were recruited at the University of California, Berkeley. Two radiologists did not finish the study, and their data were excluded. Sample size was determined based on radiologists' availability at RSNA and

was similar to previous studies on the perceptual performance of radiologists and individual differences in visual perceptual biases (Manassi et al., 2019; Manassi et al., 2021; Kosovicheva and Whitney, 2017; Z. Wang et al., 2020). Experiment procedures were approved by and conducted in accordance with the guidelines and regulations of the Institutional Review Board at University of California, Berkeley. Participants all consented to their participation in the experiment.

Stimuli and Design

Three random objects were created to simulate tumor prototypes. Between each pair of prototypes, 48 morph images were generated using FantaMorph (Abrosoft Co.). This resulted in a continuum of 147 simulated tumors in total (Figure 3.1a). In addition to the simulated tumors, 100 real mammogram images taken from The Digital Database for Screening Mammography were used in this study as background textures (Heath et al., 2001).

On each trial, one of the 147 simulated tumors was randomly chosen and presented on top of a randomly chosen real mammogram background image (see Figure 3.1b for an example trial). The simulated tumor was shown at a random angular location relative to central fixation (0.35 degrees of visual angles) in the peripheral visual field with an eccentricity of 4.4 degrees of visual angle. After 500 ms, a noise mask covered the whole screen for 1000 ms to reduce retinal afterimages. Next, one random simulated tumor image was shown at the center of the screen, and participants (trained radiologists and untrained observers) were instructed to adjust the current image to match the previously shown simulated tumor. This adjustment was performed by pressing the left and right arrow keys to move along the simulated tumor continuum.

Participants were allowed to take as much time as needed to complete this task. Once they decided on the chosen image, they confirmed their response by pressing the space bar. A brief 250 ms pause followed their response, and then the next trial began. Each participant completed 255 trials in total.

Data Analysis

For each participant, we estimated their perceptual biases with their response errors on each trial by calculating the shortest distance in morph unit on the simulated tumor continuum between the target and their response.

In order to directly compare the discriminability of the simulated tumors between radiologists and untrained observers, we calculated the just-noticeable-difference (JND) by fitting a Gaussian function on the response error frequency on individual observers, and calculated half of the distance between the 25th and 75th percentile of the cumulative Gaussian distribution that was transformed from the best-fitted Gaussian function.

Within-subject consistency in the response errors was calculated with a split-half correlation for each observer. To compensate for the lack of trials for each image, we first binned every three simulated tumors into one, so that the number of unique simulated tumors was

reduced to 49, but every binned simulated tumor had on average 5 trials of response errors. We then used a nonparametric bootstrap method to estimate split-half correlations (Efron & Tibshirani, 1994). On each iteration, for each observer and each binned simulated tumor, we randomly split the responses into two halves and calculated the mean response errors for each half (see Fig. 3.2a and 3.2d for the two randomly-split halves from all radiologists and Fig. 3.3a and 3.3c for all untrained observers). Next, the two halves were correlated and then the Pearson's r value was transformed into a Fisher z value (see Fig. 3.2b and 3.3b for the individual within-subject correlations for each radiologist and each untrained observer). We then averaged the z values from radiologists and untrained observers separately and the averaged Fisher z values from two groups were transformed back to Pearson's r values (Fisher transformations were applied for all analyses when calculating the average of correlation values). We repeated this procedure 1000 times so that we could estimate the mean within-subject correlations and 95% bootstrapped confidence intervals (CI) for radiologists and untrained observers separately (Figure 3.2c, left panel).

Between-subject consistency was calculated similarly. After splitting every observer's data into two random halves (i.e., by randomly selecting 50% of the data on each iteration), we correlated one half from one observer with one half from another observer. All pairwise correlations were averaged to estimate the between-subject consistency. By repeating the procedure 1000 times, we obtained the mean between-subject correlations and 95% bootstrapped CIs separately for radiologists and untrained observers (Figure 3.2c, right panel).

Next, we estimated the expected chance-level within and between-subject correlations by calculating permuted null distributions. On each iteration, and for each observer, we again split the response errors for each binned simulated tumor into two halves as we did in the bootstrap procedure. We then systematically shifted one half by some random units (for example, simulated tumors 1, 2, 3 might be labeled as simulated tumors 7, 8, 9), and the shifted half was correlated with another unchanged half. For within-subject correlations, the unchanged half came from the same observer. For the between-subject correlations, the unchanged half came from a different observer. The resulting correlations from individual participants (within-subject) or different pairs of participants (between-subject) were averaged together to get the permuted within-subject or between-subject correlations. This permutation method allowed us to estimate the null correlations by correlating the response errors of different stimuli with each other while at the same time preserving the relationship between similar stimuli (Monte Carlo Permutation Test, MCPT; Dwass, 1957; Edgington and Onghena, 2007; Manly, 2018). This permutation procedure was repeated 10,000 times to estimate permuted null distributions for within-subject and between-subject consistency. We did this separately for radiologists and untrained observers. The mean empirical bootstrapped correlations were then compared to their corresponding permuted null distributions to estimate the statistical significance of the mean bootstrapped within and between-subject correlations.

Internal consistency of the stimuli used in the experiment was calculated using Cronbach's alpha (Cronbach, 1951). We first binned the simulated tumors into three categories (i.e., the three prototypes). Each image was labeled as the closest simulated tumor prototype based

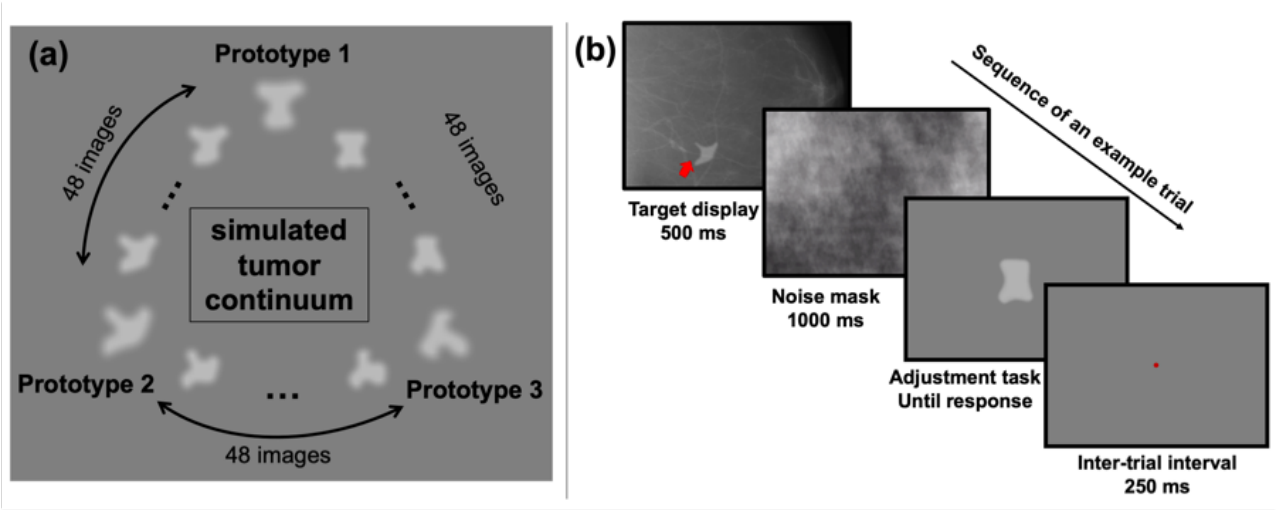

Figure 3.1: Stimuli and an example trial in Experiment 1. (a) The simulated tumor continuum was generated from three tumor-shape prototypes. Between each pair of prototype images, 48 morph images were generated, resulting in a total of 147 images. (b) On each trial, observers saw a simulated tumor target (indicated by the red arrow; note that the red arrow was never shown in actual experiment) superimposed on a real radiograph for 500 ms, which was followed by a noise mask covering the full screen for 1000 ms. Then a randomly chosen simulated tumor was shown at the center of the screen and observers pressed the left or right arrow key to adjust it and match it with the target. They confirmed their response by hitting the space bar and after a 250-ms inter-trial-interval, the next trial began.

on its distance on the simulated tumor continuum. Participants' responses (i.e., selected images) were also transformed into three categorical responses as above. We estimated Cronbach's alpha separately for radiologists and untrained observers.

Results

Our goal was to measure individual differences in radiologist perception and compare these to the individual differences found in a sample of untrained observers. In Experiment 1, we used artificial radiographs: images with controlled shapes that were presented briefly in noise (Figure 3.1a). The background noise was taken from authentic radiographs (Heath et al., 2001), and was therefore realistic. The simulated tumors, on the other hand, were intentionally artificial because we aimed at having highly controlled stimuli with solid ground truth information; this allowed us to precisely measure perceptual biases in judgment. On each trial, the clinician saw a brief image of a radiograph with a simulated tumor. The clinician was asked to find the simulated tumor in the noise image, and then to match that simulated tumor with a test stimulus in a continuous report paradigm (Figure 3.1b).

The advantage of this paradigm over categorization or forced-choice tasks is that it gives trial-wise errors and allows us to measure a complete error distribution with high-resolution information. The goal here was not to recreate a diagnostic imaging task, but to measure perceptual biases for visual stimuli that used noise backgrounds similar to those found in typical radiographs.

The results showed that the practicing radiologists are able to match the artificial tumor with a corresponding shape very accurately: mean JND was 10.5 morph unit with standard deviation 2.0 morph unit). This confirms that they were able to detect and recognize the simulated tumors. Our goal was to look for individual differences that may have been stable and consistent within a particular observer—whether there are idiosyncrasies in clinician perception. To measure this, we calculated the consistency in the observer judgments of the simulated tumors. Each simulated tumor was different, and we measured systematic errors in judgments for each specific image. Insofar as there are differences in clinician perception, they might report deviations or biases and (mis)report a simulated tumor consistently.

Figure 3.2a shows an example of an individual radiologist's biases as a function of stimulus number and all remaining radiologists are shown in Fig. 3.2d. We calculated the split half correlation within each observer (Fig. 3.2b) across all of the stimuli and found that there was a significant within-participant correlation (Fig. 3.2c, left panel, red bar; mean Pearson's $r = .37, p < .001$, permutation test). Hence, each observer had idiosyncratic biases in their perceptual reports, and those were consistent within each observer. We also calculated the between-observer correlation, using the same approach. This is the correlation between different clinicians, or how similar their residual errors were to each other; it is a measure of how much agreement there is between observers. We found that there was significantly more correlation within a given clinician than between clinicians (Fig. 3.2c, right panel, red bar; mean $r = .22$, bootstrap test, $p < .001$). This cannot be attributed to noise. Simply adding noise reduces the correlation both within and between observers; adding noise can't increase the within observer correlation. The results suggest that individual clinicians have consistent biases in their perceptual reports. The source of these biases is unclear, but they are observer-specific.

To compare this sample of clinicians with an untrained group, we collected data on the same experiment with another group of naive untrained non-clinical observers (Fig. 3.3). The observers performed the exact same continuous report matching task. Untrained observers also perceived the simulated tumors accurately (mean JND: 10.0 morph unit, standard deviation: 1.9 morph unit; see Experiment 1, Data Analysis for the estimation of JNDs) and their discriminability did not differ from that of radiologists ($t = 0.64, p = .53$). We also looked into the within-subject and between-subject consistency among untrained observers and found qualitatively similar results (Fig. 3.2c). First, there were significant individual differences in the untrained observers (Fig. 3.2c, left panel, yellow bar; mean $r = .30, p < .001$, permutation test; individual observer within-subject correlations in Fig. 3.3b). The between-subject correlation was significantly lower (Fig. 3.2c, right panel, yellow bar; mean $r = .17, p < .001$, bootstrap test). This echoes the group of radiologists: there are individual differences in simulated tumor recognition, even in untrained observers.

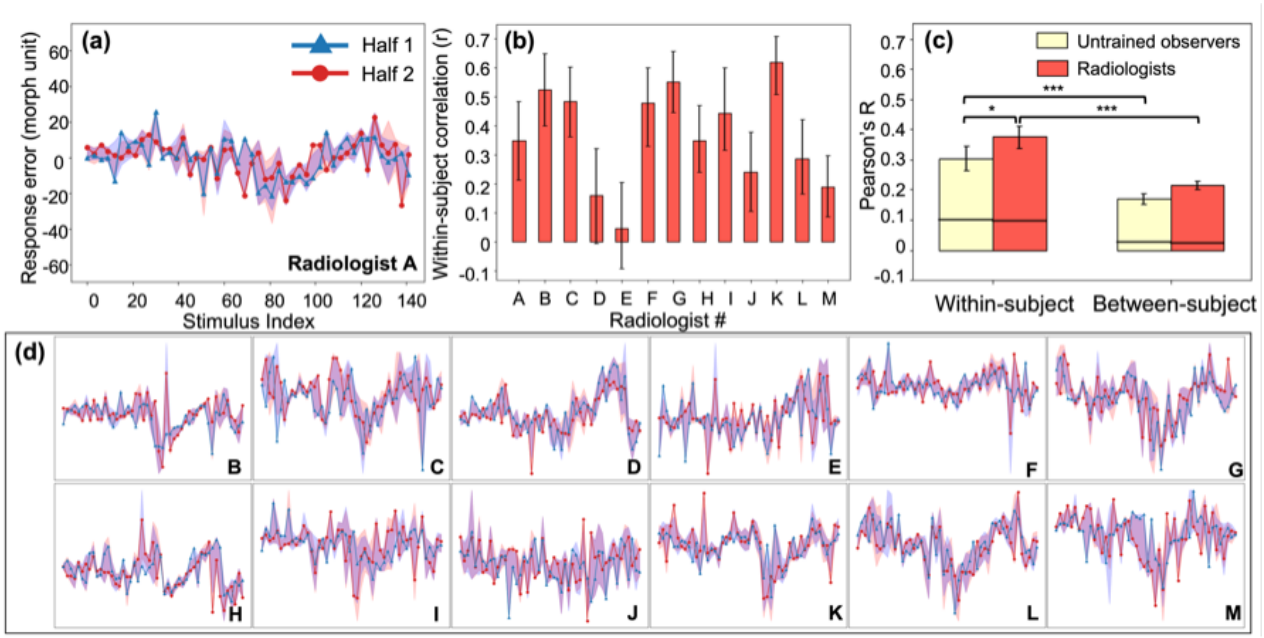

Figure 3.2: Figure 2. Experiment 1, individual differences in radiologists' perception. (a) An example radiologist's error plot. The abscissa shows the stimulus index (from Fig. 1a) and the ordinate is the observer's continuous report error. Because of the limited number of trials, data were down-sampled by binning every three neighboring simulated tumors on the continuum into one; this resulted in 49 instead of 147 data points (see Experiment 1, Data Analysis). For the purposes of visualization, we randomly split the data into two halves (red and blue curves) and plotted them separately (in the analysis this procedure was repeated, see Methods). The shaded area around the two halves represents the 95% bootstrapped confidence interval for each half. (b) The average bootstrapped split-half within-subject correlation for each radiologist. Error bars represent the 95% bootstrapped CIs for each individual radiologist. (c) Within-subject consistency and between-subject consistency were averaged within each group of observers (radiologists in red; untrained observers in yellow). For both groups, the within-subject correlations were significantly higher than between-subject correlations, and radiologists were significantly more consistent within themselves compared to untrained observers. Error bars represent the 95% bootstrapped CIs and the horizontal black lines represent the 97.5% upper bound of the permuted null distributions. *: $p < .05$, ***: $p < .001$. (d) Individual radiologist's error plots for Radiologists B-M. Strong idiosyncrasies are clear between different radiologists while at the same time there are noticeable consistencies within each radiologist, indicating stable response biases. The abscissa and ordinates for error plots in (d) are exactly the same as those in (a) so for visual simplicity, they are not labeled.

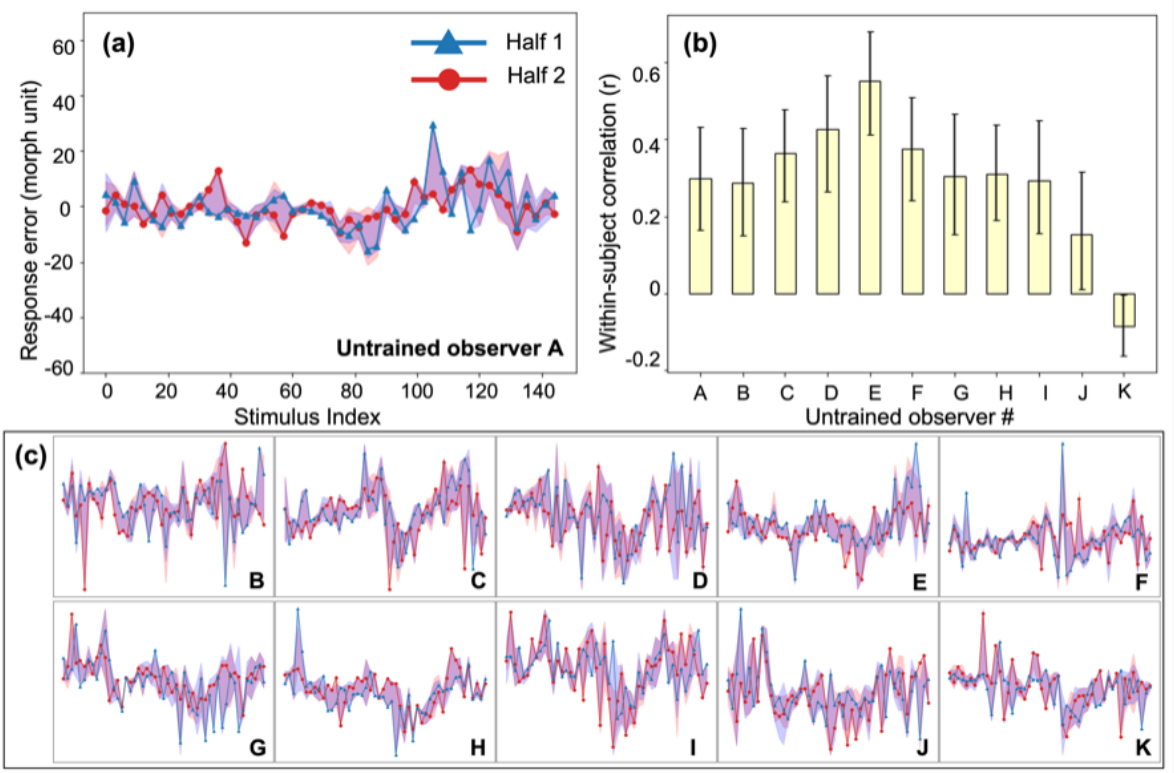

Figure 3.3: Experiment 1, individual differences in untrained observer perception of simulated tumors. (a) An example untrained observer's error plot (as in Fig. 3.2). The shaded ribbons around the two halves (red and blue data) represent the 95% bootstrapped confidence intervals for each half. (b) The bootstrapped within-subject correlation for each untrained observer (c.f., the group average in Fig. 3.2c). One observer (Observer K) showed a moderately negative within-subject correlation, but this could be due to chance. Error bars represent the 95% bootstrapped CIs. (c) Individual subject error plots for untrained observers B-K. The individual differences previously found among radiologists (Fig. 3.2) were replicated with the untrained observers. Ordinate and abscissa for each error plot in (c) are identical to those in panel (a).

There are, however, several differences between the radiologists and untrained observers that are worth noting. First, the within-observer correlation was higher for the radiologist group than for the untrained observers ($p < .05$, bootstrap test). Clinicians are more consistent in their observer-specific biases than untrained observers. Given that clinicians and untrained observers do not differ significantly in their perceptual sensitivity measured by JNDs, this result echoes our hypothesis that idiosyncratic perceptual biases could be observed even without differences in overall perceptual sensitivity. Second, the between-subject correlation was not 0 in either group ($ps < .001$, permutation test). There are therefore some consistencies between observers in how these stimuli are judged. The individual differences, however, significantly outweighed the commonality, since the within-subject correlations of both groups were significantly higher than the between-subject correlations ($ps < .001$). Together, the results in Experiment 1 showed that radiologists and untrained observers both demonstrated strong individual differences in their perceptual biases towards different simulated tumors in a shape matching task, and radiologists tend to have higher consistency in their own biases.

However, several questions were still left unanswered. First, in Experiment 1, although we used real mammograms as backgrounds, the simulated "tumors" were clearly artificial and different from real tumor shapes (Fig. 3.1a). It is unclear whether radiologists would show any idiosyncratic perceptual biases on real or very-close-to-real radiographs. Second, it remains unknown whether these perceptual biases can be observed in perceptual tasks other than continuous report shape matching. Third, we wondered whether similar individual differences would still exist for medical images other than mammograms. Therefore, to further explore these questions, we conducted a second experiment.

Experiment 2

Raw data for Experiment 2 were obtained from a published study (Ren et al., 2021).

Methods

Participants

Seven trained radiologists (3 female, 4 male, age: 28-40 years) and five untrained observers (3 female, 2 male, age: 23-25 years) were recruited in the experiment. Sample size was determined based on previous studies on the perceptual performance of radiologists and individual differences in visual perceptual biases (Manassi et al., 2019; Manassi et al., 2021; Kosovicheva and Whitney, 2017; Z. Wang et al., 2020). Experiment procedures were approved by and conducted in accordance with the guidelines and regulations of the Institutional Review Board at University of California, Berkeley. Participants all consented to their participation in the experiment.

Stimuli and Design

Fifty CT lesion images were randomly sampled from the DeepLesion Dataset (Yan et al., 2018), and fifty simulated lesion images were generated through the Generative Adversarial Networks (GAN) trained with 20,000 real CT lesion images from the DeepLesion Dataset (Ren et al., 2021). This resulted in a total of 100 images (see Figure 3.4a for examples). According to Yan and colleagues (2018), there were multiple types of lesions in the DeepLesion Dataset, including lung nodules, liver tumors, enlarged lymph nodes, and so on, and images included both chest and abdomen CT images.

Both radiologists and untrained observers were recruited to perform an image rating task (Figure 3.4b). On each trial, one of the 100 images was pseudo-randomly chosen to present to the observers and it remained on the screen for at most five seconds. Observers were instructed to rate the realness of the image on a continuous scale ranging from 0 to 10 (0: fake, 10: real). Participants could respond at any point during image presentation, or they could take as much time as necessary after the image offset. The next trial started immediately after their response. Each image was shown exactly once in these 100 trials. Both radiologists and untrained observers were informed that the stimuli shown in the experiment were composed of 50% real images and 50% GAN-generated images.

To estimate test-retest reliability, 20 real images and 20 GAN-simulated images were randomly chosen from the aforementioned 100 images. These 40 images were randomly inserted in the previous 100 image list and were presented in the same manner. Thus, there were in total 140 trials for each participant.

Data Analysis

Due to a technical problem during image display, one of the forty repeated images failed to show up for some participants, so only the ratings for 39 out of the 40 repeated images were used (39 initial ratings and 39 retest ratings) in all following analyses.

We recognized that the raw ratings could be influenced by participants' extreme response tendencies. For example, some might tend to give higher ratings across all images while some may rate lower. Throughout the manuscript we refer to these types of response tendencies as "response propensities", to avoid confusion with other terms like response bias, that can mean different things in different circumstances. To reduce the effect of response propensities, for each participant, we first normalized their raw ratings by rescaling them to range from 0 to 10 using the equation below (X is the raw ratings from one participant, X_{min} is the minimum of this participant's raw ratings and X_{max} is the maximum).

$$X_new = (10 * (X - X_min))/((X_max - X_min))$$

After normalization, for each participant, we again used response errors as a proxy for perceptual biases. We estimated their response errors by calculating the absolute difference between the normalized ratings and the corresponding ground truth of each image (0 for GAN-generated images and 10 for real CT images). Then, similar to Experiment 1 (see

Data Analysis), we estimated the within-subject and between-subject consistency of their response errors. Within-subject consistency was estimated by the average test-retest reliability among participants (i.e., correlating the response errors from the 39 initial trials and the 39 retest trials). Figure 3.4c shows the individual split-half within-subject correlations for each radiologist and each untrained observer. Between-subject consistency was estimated as the average pairwise correlations among participants. This was calculated separately for radiologists and untrained observers.

Bootstrap distributions of the within and between-subject correlations were estimated to test whether the average correlations were simply driven by extreme observer(s). For within-subject correlations, on each iteration, we randomly sampled seven radiologists and five untrained observers with replacement, calculated each observer's within-subject correlation and then averaged the correlations through Fisher transformation (Figure 3.4d, left panel). For between-subject correlations, on each iteration, we sampled the same number of pairs of subjects from all possible pairs of subjects with replacement, calculated all pairwise correlations for the sample and the between-subject correlation was estimated as the mean of all pairwise correlations (Figure 3.4d, right panel). These procedures were repeated 1000 times to estimate the 95% within-subject and between-subject bootstrapped CIs for radiologists and untrained observers separately.

To examine the expected correlations by chance, permuted null distributions were also calculated in Experiment 2 by shuffling the image labels for the initial trials and the retest trials, so that the response error for one image might then be treated as the response error for another. Null distributions were separately calculated for radiologists and untrained observers. On each iteration, the shuffled response errors from initial trials and retest trials from the same observer were correlated to obtain a null estimate of within-subject consistency. This was done for every observer (every radiologist and every untrained observer), and the set of pairwise correlations were averaged across the group of observers to create one null sample. This procedure was repeated 10,000 times to create a null within-subject distribution. To create between-subject null distributions, we calculated all pairwise correlations between the shuffled initial response errors from one observer and the shuffled retest response errors from another observer. This was repeated 10,000 times to generate a between-subject null distribution. Permuted null distributions were calculated separately for radiologists and untrained observers, and, from these, 95% permuted confidence intervals (CIs) were estimated.

Results

In Experiment 2, we tested whether idiosyncratic perceptual biases can be observed with images from a different modality (CT images), a very different perceptual task and highly realistic GAN-generated images. Generative Adversarial Networks (GAN) were trained by Ren and colleagues (2021) with real CT lesion images taken from the DeepLesion Dataset (Yan et al., 2018), and then the GAN model was used to generate artificial lesion images. Observers were recruited to perform an image discrimination task including 50 real lesion

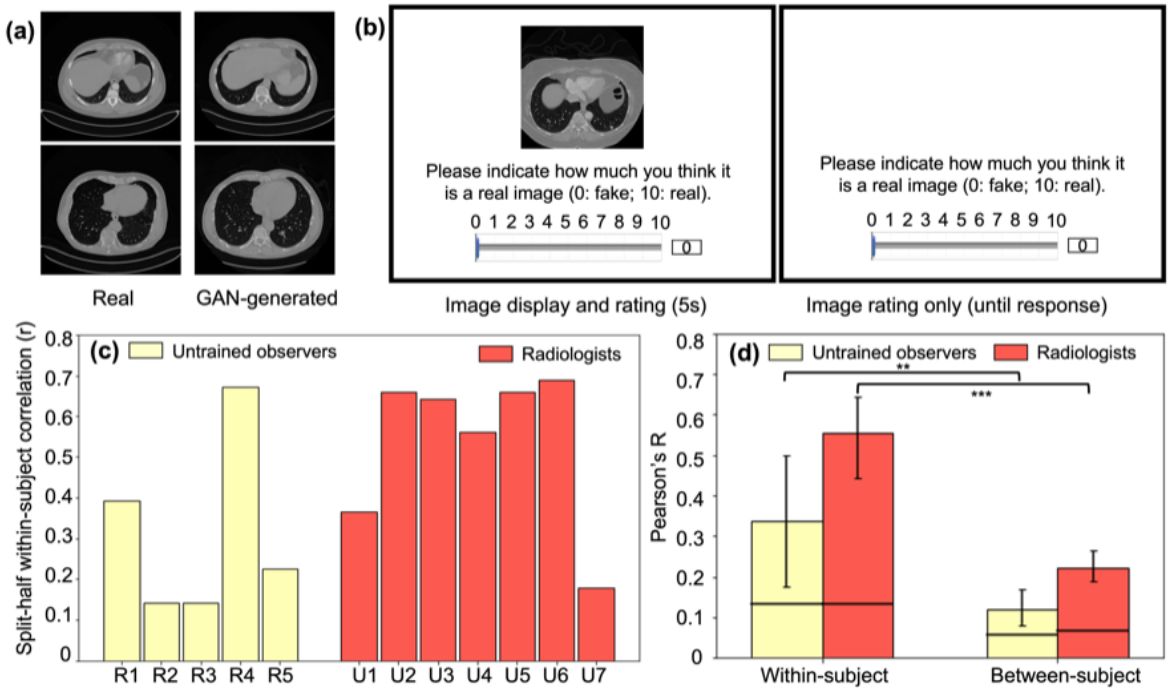

Figure 3.4: Experiment 2 example stimuli and results. (a) Examples of real (left column) and GAN-generated (right column) CT lesion images used in the experiment. Visually and qualitatively the images are similar, and a previous study showed that these GAN-generated images were visually metameric: they were indistinguishably realistic to both radiologists and untrained observers (Ren et al., 2021). (b) An example trial in Experiment 2. On each trial, either a real CT lesion image or a GAN-generated image was shown on the screen for at most 5 seconds. Participants could rate the realness of the image from 0 to 10 (0: fake, 10: real) anytime during the image presentation or after the disappearance of the image (self-paced). The next trial began after the rating was made. (c) Within-subject consistency for each radiologist and untrained observer, calculated from the split-half correlation of each observer's response errors of the same images. (d) Within-subject correlations significantly exceeded between-subject correlations for both groups ($p < .01$ for untrained observers and $p < .001$ for radiologists), which replicated the individual differences found in Experiment 1 and extended the results to a new task, a new type of medical image, and with more realistic stimuli. Error bars represent the 95% bootstrapped CIs and the horizontal lines represent the 97.5% upper bounds of the permuted null distribution. **: $p < .01$, ***: $p < .001$.

images and 50 GAN-generated images (see Figure 3.4a for example stimuli), in which both radiologists and untrained observers rated how realistic each image appeared. Among these 100 images, 20 real images and 20 artificial images were repeated twice so that we could estimate the within-subject consistency. Figure 3.4b shows an example trial.

In general, the GAN-generated images were indistinguishable from real lesion images for both untrained observers and radiologists (mean d' values: 0.18 and 0.27 respectively) and there was no significant difference between different groups of participants ($t = 0.4, p > .5$). This suggested that the artificial GAN-generated images were highly realistic and even experts with training could not distinguish them effectively (Ren et al., 2021). This general lack of sensitivity, however, does not preclude individual biases in the discrimination of the CT lesion images.

The goal of the following analysis was to measure whether there are systematic and idiosyncratic stimulus-specific biases in the perception of CT lesions by radiologists and untrained observers. As in the first experiment, we measured within and between subject consistency of the perceptual judgments. Since raw ratings may be subject to observers' response propensities, we normalized the ratings for each observer and then calculated response errors to get a more accurate estimate of their perceptual biases based on the normalized ratings (see Experiment 2, Data Analysis). Figure 3.4c shows each observer's within-subject consistency and the mean within-subject and between-subject consistency is shown in Figure 3.4d. We again found a significantly higher within-subject consistency (Fig. 3.4d, left panel) in both radiologists (Pearson's $r = .56$) and untrained observers (Pearson's $r = .34$) compared to their corresponding between-subject consistencies (Fig. 3.4d, right panel; $r = .22$ and $r = .12$, bootstrap test, $p < .001$ and $p < .01$, for radiologists and untrained observers respectively). This replicated the findings in Experiment 1, indicating that each radiologist and each untrained observer have their own unique biases in the perception of medical images that cannot be explained by shared biases among observers. We again found that between-subject correlations were significantly higher than the permuted null correlations for both groups of observers (permutation test, $p < .001$ and $p < .01$ for radiologists and untrained observers respectively), suggesting that observers do share some of their biases. This could be due to some textures or features of the images that commonly influenced the observers' discrimination of the real or fake CT lesion images.

Taken together, in Experiment 1 we found idiosyncratic biases in radiologists when they made perceptual judgments about artificial simulated tumor shapes, and their biases were stronger compared to untrained observers. Experiment 2 further extended these results and demonstrated strong individual variations in radiologists' biases in the perceived realness of GAN-generated and real CT lesion images, suggesting that idiosyncratic perceptual biases among radiologists are not tied to a specific type of medical images or tasks, but rather they can be generally observed among different modalities of medical images and different tasks. These individual observer specific biases are found even without significant difference between observers' perceptual sensitivity measured by d'.

Discussion

We found significant individual differences in radiologists' perceptual biases. Experiment 1 showed that each radiologist demonstrates unique perceptual biases towards simulated tumors in a shape matching task, and their own internal biases were even more consistent than untrained observers (Fig. 3.2c). Experiment 2 replicated and extended the results by again showing individual differences in radiologists when they perceived GAN-generated and real CT lesion images in an image discrimination task (Fig. 3.4d). These individual differences were not simply induced by task or stimuli since they were found across different radiologists, different modalities of medical images, and different tasks. Thus, we propose that the individual differences in radiologist perception may arise at least in part because of distortions in perceptual judgments at the level of specific clinicians. These kinds of individual differences in basic perceptual bias could, in turn, potentially influence the performance of radiologists in diagnostic practice.

What are the potential mechanisms underlying these idiosyncratic perceptual biases among radiologists? One possibility is that different radiologists may have different perceptual templates or perceptual representations of the tumor or of medical images in different modalities, analogous to studies showing that human observers have different perceptual templates of faces (e.g., Dotsch and Todorov, 2012; Moon et al., 2020). Differences in these templates could be associated with different biases, and could arise as a natural consequence of their intensive training and years of experience (Imhoff et al., 2011; Jack et al., 2012; Soto, 2019). Previous research has supported that attention towards perceptual stimuli can be guided differently according to the perceptual templates or mental representations (e.g., Griffin and Nobre, 2003; Olivers et al., 2011), so the variations in observers' mental representations could potentially direct their attention to different parts of the simulated tumors or the radiograph background and thus lead to idiosyncratic perceptual biases. It remains unknown if this is the case, but it is worth pursuing in future research.

Another possible explanation is natural statistics. Radiologists may not have literal "templates," but may have some priors or learned distributions of the statistics in medical images, similar to how human observers represent the statistics of scenes (Torralba and Oliva, 2003; Stansbury et al., 2013). These priors may include low-level information such as luminance and contrast, but may also contain higher-level, multi-dimensional representations of textures. These kinds of image statistics could underlie perception of gist in medical images (Evans et al., 2019). Sensitivity to this information can be shaped specifically by the natural statistics of medical images that clinicians are exposed to during their medical training and diagnostic practice, and it can also come from other non-specific visual images or perceptual experiences in their everyday life. This could explain why the GAN-generated medical images may have confused the experts in Experiment 2, since the image statistics would have been learned and captured by the GAN. Our current study cannot pinpoint the underlying mechanism responsible for idiosyncratic perceptual biases, but image statistics or templates (or both) could be involved. It would be valuable to explore this in future research.

There are several concerns raised by our results that we address here. First, it might be argued that stronger idiosyncratic biases in the radiologists in Experiment 1 could simply result from the radiologist group being more attentive to the task or lapsing less frequently. In principle, that may explain the higher within-subject correlation as well as the higher between-subject correlation in Fig. 3.2c. Although this is possible, this does not seem likely, as the overall discriminative ability was fairly similar between the two groups and the just noticeable difference was comparable and not significantly different for both the clinician group and the naïve untrained group (see Experiment 1 results). Hence, the stronger within-subject correlations—stronger individual differences within the clinicians—did not simply arise because of attentiveness. On the other hand, the stronger individual differences in the clinician group could arise, in part, from the unique and lengthy training that the clinicians receive, or the practice that they have had in related types of perceptual tasks.

One limitation is that the task we used in the Experiment 1 was not realistic and arguably was not representative or typical of a radiologist's task because radiologists mostly perform detection or categorization tasks in their everyday routine, while our task was a continuous report adjustment method. However, the adjustment task can be more advantageous than detection or categorization tasks since it can measure the subjective perceptual representations and criterion of the observers (Pelli & Farell, 1995), it provides very fine-grained information about errors, and it provides critical behavioral insights for understanding the perceptual biases in medical image perception. Moreover, there is no evidence that we know of that continuous report psychophysical measures systematically misrepresent recognition processes (e.g., Prinzmetal et al., 1998; see Stevens, 1958; Gescheider, 2013 for reviews). Nevertheless, future studies could extend our results by testing radiologists with our controlled realistic stimuli in a task that is more similar to those in clinical practice.

Another related concern is that the task in Experiment 1 may be unrealistic because it required a variety of perceptual and memory related skills. Observers (naïve observers or skilled radiologists) were asked to detect and recognize a simulated tumor. During this task, they had to hold information in visual short-term memory and subsequently match a stimulus to what was previously seen. This is indeed broader in scope than a traditional forced-choice paradigm. Nevertheless, the detection, recognition, and visual short term memory processes involved in our task are the kinds of abilities that are used by clinicians on a daily basis. Multitasking is not uncommon for radiologists in realistic settings; they often have multiple screens and multiple radiographs; they gaze between different regions of the visual field and integrate information separately from multiple radiographs and files; radiologists often need to hold in short term memory information about the patient, diagnostic history, etiology, referring physician, etc.; and, they may be interrupted mid-diagnosis by the phone, noise, and other realistic factors (see Kansagra et al., 2016 for a review). Moreover, visual short term memory processes work even at the shortest time scales (e.g., even across saccades; Irwin, 1991). In other words, the complexity of the radiologist's task goes well beyond a simple instantaneous forced-choice.

Our experiment does not capture the full complexity of the radiologist's family of tasks, but the basic processes it taps are highly relevant to those used by radiologists. The re-

sults reinforce this: radiologists had higher within-subject consistency than the untrained observers. This suggests that individual radiologists have more consistent and systematic biases in this simulated tumor matching task compared to untrained observers, indicating that their expertise or experience is in fact reflected in this task. Although radiologists and untrained observers had similar sensitivity, as measured by JNDs, this is not surprising since previous studies have found that untrained, naïve observers can perform significantly better than chance in the Vanderbilt Chest Radiograph Test Sunday et al., 2017, and other studies found that MDs are not always more sensitive than untrained participants in medical image perception tasks, and sometimes they even have lower sensitivity compared to less experienced observers (e.g., Wolfe, 2022). The similar sensitivity in MDs and untrained observers in our experiment could be due to a ceiling effect in our data, but the fact that the consistency in reports is higher for MDs suggests that they do, in a sense, perform the task better than untrained observers.

One way to address this question about ecological validity is to test whether our results extend to other tasks, especially involving real medical images and stepping beyond the artificial radiographs. Therefore, we analyzed data from a second experiment that used realistic CT lesion images. Although this is a different area of medical image perception, we hypothesized that idiosyncratic perceptual biases can be observed across domains and should not be limited to any particular modality, stimulus, or task. In the second experiment, we used real CT lesion images combined with artificial but realistic lesion stimuli created by Ren and colleagues (2021). These stimuli (see Fig. 3.4a for example) are different than artificial stimuli in Experiment 1 (Fig. 3.1a) because they were highly realistic, even metameric (completely confusable) with real lesions (Ren et al., 2021). Using those realistic images, we found that both radiologists and untrained observers showed clear individual differences in their perception of the real or fake CT lesion images, which extended our previous findings from a matching task to a real/fake rating task and from artificial shapes to highly realistic CT images. The results from Experiment 2 further supported our hypothesis that observer-specific perceptual biases are not domain modality or task-specific. Rather, they are likely a ubiquitous effect in realistic medical imaging tasks with implications across domains. Therefore, though our tasks might not be the most realistic or cannot be directly linked with diagnostic performance of radiologists, these compelling results clearly demonstrate that even well-trained radiologists can have idiosyncratic and stimulus-specific perceptual biases with medical images under different task settings.

One might still be concerned about the internal consistency for these idiosyncratic biases. Using the split-half Pearson's correlation, we found that radiologists had an internal reliability of 0.37 (Experiment 1) and 0.42 (Experiment 2). While this may seem somewhat low, it is significantly higher compared to what was expected by chance (i.e., the permuted null distributions) and it may appear low only because our stimuli were numerous and very finely spaced. In order to compare our results with previous published studies, in Experiment 1, we dummy-coded the data into binned categories (like a three-alternative-forced-choice, 3AFC, classification task, see Experiment 1, Data Analysis for details) and the Cronbach alpha rises substantially (alpha = 0.85 for radiologists), and in Experiment 2, the Cronbach

alpha for radiologists was 0.95, which are indeed comparable to that reported in a previous study on individual difference in a radiograph-related task (Sunday et al., 2017). This is unsurprising because noise at the individual stimulus level is averaged out and what remains is a less noisy estimate of the more substantial individual differences.

The between-observer consistency is typically the focus of most medical image perception research (see Donovan et al., 2017 for a review; Elmore et al., 1994; Feldman et al., 1995; Beam et al., 1996; Tan et al., 2006; Lazarus et al., 2006; Elmore et al., 2009; Donovan and Litchfield, 2013; Sonn et al., 2019; Pickersgill et al., 2019). Recently, a study by Sunday and colleagues (2017) explored the internal within-observer consistency in medical image perception. Our results complement this from a different perspective, showing that it is equally or even more important to measure individual differences in each radiologically-relevant task by measuring both the within-subject and between-subject consistency in radiologists. Our results also go a step further to compare the individual differences in radiologists with untrained non-clinical observers, and they provide evidence for a stronger idiosyncrasy in radiologists' perception of artificial and realistic medical images across domains, which was not clear in previous studies on individual differences in radiologists. The stronger within-subject consistency compared to between-subject consistency also provides a direct insight about the relative importance of the individual perceptual variations and shared biases among observers: individual differences are substantial, and can even swamp the between-subject similarities.

The scarcity of the expert radiologist pool undoubtedly limited the number of available observers we were able to test. Although this is a limit in group-wide analyses, we analyzed every individual observer and measured trial-wise effects within each observer. In fact, even when sample size was limited, past research has been able to demonstrate strong and consistent idiosyncratic visual perceptual biases towards object location, size, motion and face perception with the help of psychophysics (Afraz et al., 2010; Schütz, 2014; Z. Wang et al., 2020). Our results aligned with these previous findings and provided new insights of the prevalent idiosyncratic perceptual biases that can be found with medical images and among well-trained radiologist experts.

There are several implications of the findings reported here. First, clinicians vary in their perceptual abilities. Although this is not at all surprising, the stimulus-specific way in which clinicians vary in the perceptual biases is novel. Second, we found that individual differences are not washed out by training. To address this, we performed a Fisher's combined probability test (Fisher, 1992; Fischer et al., 2011; Rosenthal, 1978), which combines the statistical results from both Experiment 1 and Experiment 2 in a type of "mini meta-analysis" (Goh et al., 2016). We found that, across the experiments, there is a significantly higher within-subject consistency for radiologists compared to untrained observers ($\chi^2_2 = 12.9, p < .005$). That is to say, counterintuitively, some biases may get stronger with training, leading to more stable individual differences within radiologists compared to untrained observers. Combined with the fact that radiologists and untrained observers were not significantly different in terms of perceptual sensitivities (measured by JNDs in Experiment 1 and d' in Experiment 2), this result again echoes our hypothesis that variations in perceptual biases

could exist even without overall differences in perceptual sensitivity. Third, our results show that even untrained observers bring with them individual biases and idiosyncrasies in their perceptual judgments. Fourth, idiosyncratic distortions were found across two different domains, two different modalities, and two different imaging techniques (see Experiment 1 vs Experiment 2).

More importantly, the fact that there are individual differences between observers could have critical implications for diagnostic medical imaging. For example, in some countries, it is common practice to have multiple readers rate or diagnose radiographs (Australia, 2002; Yankaskas et al., 2004; Amendoeira et al., 2013). Given that much of the variance in observer judgments is attributable to the individual observer themselves, it may directly influence the employment or selection of the pairs of observers: two observers that have similar individual differences may perform more poorly (since the biases could potentially exaggerate after combining) than two readers who have more independent individual differences. Individual differences that are more independent will tend to cancel out and thus lead to more accurate medical image perception (Van Such et al., 2017; Corbett and Munneke, 2018; Taylor-Phillips and Stinton, 2020). Thus, our results may suggest the importance of measuring perceptual biases in radiologists before grouping them into pairs. We believe this could be a valuable strategy in paired reading, as it goes well beyond simply relying on radiologists' diagnostic accuracy (e.g., Brennan et al., 2019).

Another important implication of our results is that different clinician observers may show different native ability in particular specialties or even different imaging modalities. Returning briefly to the face recognition literature, the individual differences in face recognition arise because of many factors including age and experience, but also genetic differences (Wilmer et al., 2010; Shakeshaft and Plomin, 2015). Some observers are simply genetically predisposed to be more sensitive to faces. The same may be true in medical image perception. Although we do not know what proportion of the individual differences are accounted for by genetic factors, this will be an important future area of research. Whether or not there is a substantial genetic contribution, the individual differences in clinician perception can also be measured, selected, and trained. Some previous work has already started to address this with mostly focusing on perceptual sensitivity (Corry, 2011; Birchall, 2015; Langlois et al., 2015; Sunday, Donnelly, & Gauthier, 2017, 2018; see Waite et al., 2019 for a review). The individual differences reported here also echo those found in other domains of perception research (e.g., Wilmer et al., 2010; Schütz, 2014; Wexler et al., 2015; Grzeczkowski et al., 2017; Z. Wang et al., 2020; Canas-Bajo and Whitney, 2020), and raise the possibility that idiosyncratic distortions in clinician perception may be widespread and extend across different domains.

Our findings provide a new insight about the individual differences that exist in the perceptual judgments of professional radiologists: apart from perceptual sensitivity, which has been proposed and investigated extensively in the past, there may actually be idiosyncratic and systematic biases in their perceptual judgments. Understanding these idiosyncratic perceptual biases could be critically important for a variety of reasons, including training, career selection, bias compensation, and employing paired readers in the field of medical imaging.

At an even broader level, it is worth noting that individual differences in observer perception could have important consequences in many fields beyond medicine. For example, in TSA screeners, professional drivers, airline pilots, radar operators, and in many other fields where single observers are relied on for life-altering perceptual decisions.

Together, in Chapter 2 and Chapter 3, we have shown that individual differences in perceptual biases can be widely found in a variety of tasks and even among experts. These results together can have pragmatic implications: employing pairs of readers is a powerful and commonly used approach to improve the accuracy of many occupations including radiology, pathology, TSA airport screening, radar operators, and many other fields. Between each reader, their perceptual biases could vary, leading to some pairs that are more uncorrelated and independent with each other, while some are more associated and less independent. Can pairing individual observers based on their idiosyncratic perceptual biases improve their collective decisions? The goal of Chapter 4 is to test this possibility.

Chapter 4

Diversity matters: improve collective decisions made by individuals with idiosyncratic biases

Human perceptual decisions are critically important. They can be socially significant, life-changing or even catastrophic if wrong, as demonstrated in occupations such as radiologists, TSA screeners, trial judges, sports referees, and pilots. Although these crucial judgments are made across a range of typical situations ranging from hiring and academic grading to driving, they often hinge on an individual's expertise. An inaccurate medical diagnosis from a radiologist may result in the loss of a patient's life; a detection failure of a TSA screener can lead to tragedy, and an inattentive driver can result in a disastrous accident. Therefore, the expertise of individual humans matters.

Research has widely shown that individuals are substantially different in their expertise, especially how much sensitivity could vary from one person to another. Starting from fundamental visual acuity such as contrast sensitivity, all the way to higher-level face recognition ability, the variations in human sensitivity has been well documented (e.g. Peterzell et al., 1995; Wilmer et al., 2010; R. Wang et al., 2012; Grzeczkowski et al., 2017), and they can distinguish experts from amateurs. However, is it just about good or bad?

Take the Olympic figure skating competition for an example. Skaters' performance is judged based on the difficulty, execution, timing, presentation and skating skills (International Skating Union, 2022), and these ratings are largely dependent on the judges' perceptual decisions, especially in subjective items such as presentation and skating skills. Every judge is an expert, which means that they have high sensitivity or accuracy in judging skaters' performance. However, every judge is also characterized with different biases in their ratings (Ste-Marie and Lee, 1991; Whissell et al., 1993; Findlay and Ste-Marie, 2004; Zitzewitz, 2006; Linacre, 2009). For example, studies have shown that some judges may tend to give higher scores to skaters from countries that are federated with their home country (Zitzewitz, 2006). In the context of Olympic games, even just a 0.1 deviation in the ratings could be the determining factor for a gold medal. If, by chance, several judges with similar biases

were assigned to assess the same performance, their biases could accumulate, leading to a highly-biased and inaccurate rating.

Another relevant and life-altering situation is radiological diagnosis. In many European countries such as the United Kingdom, a pair of radiologists evaluate the same radiograph and their judgements are integrated together to form the final diagnostic decisions (Wilson & Liston, 2011). Intuitively, patients may prefer two radiologists to give similar judgments. However, as revealed by recent research, each radiologist is characterized with unique perceptual biases towards medical images (Manassi et al., 2019; Manassi et al., 2021; Wolfe, 2022; Wang, in press). Given that, is it truly beneficial for the patients to have their radiograph examined by two radiologists who potentially have similar biases?

Under most circumstances like those above, where decisions are made by multiple individuals, random selection or employment based on availability was the most common practice to form the decision-making committee (e.g., Bruce, 1935; Metz and Shen, 1992; Wallsten et al., 1997; Armstrong, 2001; Surowiecki, 2004; Jiang et al., 2006; Juni and Eckstein, 2017). Given that observers tend to be characterized with unique biases, will selecting Olympic judges or pairing radiologists who have similar, unrelated or even opposite biases help them achieve more accurate decisions? Even just a 0.01% improvement in their accuracy could be highly consequential, so it is worthy of the effort to explore potential strategies to optimize their collective decisions.

In this series of studies, we aimed to examine this possibility by investigating the effect of variations in human perceptual biases on the collective decisions made by multiple observers. We tested observers in three distinct tasks: object localization, object recognition, and a subjective emotion tracking task. To foreshadow our results, we revealed that human biases played an important role in the optimization of the collective decision. Throughout all three extensively-different tasks, we observed that pairing individuals whose biases were most independent (i.e., uncorrelated) with each other can significantly outperform pairs that are randomly chosen. This finding was replicated and extended from low-level localization task to higher-level object recognition and even an emotion tracking, where there was no objective ground truth and closer to real-world situations such as sports and academic ratings or other social contexts such as trial judgement. Our results highlight the importance of understanding and measuring biases of human in both research and also in the practical and professional settings that involve the evaluation of collective decision, which have far-reaching implications for many critical situations or occupations such as employing a group of Olympic judges or trial judges, as well as selecting pairs of radiologists, TSA screeners or aircraft pilots.

Experiment 1

Method

Participants

Fifty observers (30 females, age range: 18 - 29) participated in this experiment. All subjects reported to have normal or corrected-to-normal vision. Experiment procedures were approved by and conducted in accordance with the guidelines and regulations of the Institutional Review Board at University of California, Berkeley. Participants were provided with informed consent approved by the Institutional Review Board at University of California, Berkeley.

Stimuli, Procedure and Apparatus

All visual stimuli were presented on a 19-inch gamma-corrected Dell P991 CRT monitor (Dell, Round Rock, TX; 1024 $\times$ 768 pixels resolution, 100 Hz refresh rate), black taped to reduce the influence of reference frame. Stimuli and the experiment were programmed using MATLAB (The MathWorks, Natick, MA) and Psychophysics Toolbox (Version 3, Brainard and Vision, 1997) and run on an Apple Macintosh computer (Apple Inc., Cupertino, CA). Observers viewed the stimuli binocularly at a distance of 40 cm fixed by a chin rest.

Observers were instructed to indicate the location of a noise patch presented on the screen. The noise patch was composed of 0.1 $\times$ 0.1 degrees of visual angle (d.v.a.) squares and each square was randomly chosen to be black ($<$ .001 cd/m^2, measured by Minolta LS110 Luminance Meter) or white (92.6 cd/m^2). Then the squares were enveloped within a two-dimensional Gaussian contrast aperture (standard deviation: 0.75) visible within a circular aperture (diameter: 2.44 d.v.a).

Observers were instructed to maintain fixation throughout the experiment. On each trial, a black fixation dot (diameter: 0.3 d.v.a.) was presented for at the center of a gray (48.3 cd/m^2) background. After 1000 milliseconds (ms), a noise patch was shown on one of the 50 predefined locations for 50 ms (Figure 4.1a). These 50 locations were randomly selected using stratified sampling technique to ensure that they covered 2 to 10 d.v.a. in eccentricity and spread across 0° to 360° in angular positions and the same 50 locations were used for all observers. Once the noise patch disappeared, the central fixation dot turned dark gray (30.4 cd/m^2) and after 500 ms, a white response dot (diameter: 0.45 d.v.a.) appeared at the center of the screen, and observers could freely adjust the location of the response dot using the mouse to match the central location of the noise patch (Figure 4.1b). Observers were provided with unlimited time to adjust the response dot and they were instructed to confirm their location adjustment with a left click. The center of the white response dot at the time of the left click was recorded as the reported location and a new trial began afterwards. Observers were instructed to maintain fixated even when they adjusted the response dot.

There were a total of 300 trials where each location was repeatedly tested for 6 times. The order of all trials was randomized for each observer.

Data Analysis

The response dot locations reported by the observer were first transformed into polar coordinates, and then the response errors were calculated as the angular difference between the actual noise patch center location and the reported location. Aligned with previous studies, the response errors were used to estimate observers' localization biases (Kosovicheva and Whitney, 2017; Z. Wang et al., 2020).

Given that each location was repeatedly tested for 6 times for every observer, we randomly separated each observer's responses into two halves and each contained the response errors from the 3 repetitions of all locations. One half was used to calculate the similarity between different observers' localization biases, and the other half served as a cross validation to test whether averaging observers based on their localization biases can result in better localization performance.

Pairwise Pearson's correlation was calculated based on the response errors from one of the two halves of the data to estimate the similarity across observers and the correlation matrix was plotted in Figure 4.1c. These pairwise correlations were then used to determine how the fifty observers can be grouped into 25 pairs. There were four types of pairing techniques including random pairs and three other special pairs: most positively-correlated pairs, independent pairs and opponent pairs. The most positively-correlated pairs were selected by maximizing the average pairwise correlation between the 25 pairs, while the opponent pairs should minimize the average pairwise correlation, or as negative as possible. For the independent pairs, observers were grouped together so that the average absolute pairwise correlation was closest to zero (i.e., uncorrelated). Maximum weight matching algorithm (Galil, 1986) was employed to search for the most positively-correlated pairs, opponent pairs and independent pairs. Random pairs were selected by randomly dividing the 50 observers into 25 pairs.

The remaining half of the data was then used to cross validate the performance of different pairs. For the most positively-correlated pairs, we averaged the responses from each pair of observers at every location, and then calculated the Pearson's correlation between the average reported angular locations and the actual locations and then converting the correlation r values to z values through fisher transformation. Then the z values from all 25 pairs were averaged together as the average localization performance of the positively-correlated pairs (Figure 4.1d, orange line). Same procedures were applied on opponent pairs (Figure 4.1d, blue line) and independent pairs (Figure 4.1d, red line). To estimate the performance distribution of random pairs, on every iteration, observers were randomly grouped into pairs of two and the fisher z values calculated for each pair were averaged. This procedure was repeated for 10, 000 times and yielded a permutation distribution of the random pairing performance (see Figure 4.1d). The special pairs' results were subsequently compared to the random pair distribution to examine our hypothesis that pairing observers based on their biases can benefit the wisdom-of-the-crowd. We also averaged single observer's localization performance and compared it to the random pair distribution, which allowed us to test whether pairing and averaging observers can outperform single observers (i.e., wisdom of the

crowd).

Results

Results revealed that single observers' localization performance was significantly worse than the average combined response from grouping individuals into random pairs ($p < .001$, permutation test using a Bonferroni-corrected alpha, $\alpha_B = .0125$; same below), which replicated the wisdom-of-the-crowd that has been reported by many past studies (e.g., Bruce, 1935; Metz and Shen, 1992; Wallsten et al., 1997; Armstrong, 2001; Surowiecki, 2004; Jiang et al., 2006; Juni and Eckstein, 2017). Furthermore, we found that if observers were paired so that their biases were most similar to each other, their overall localization performance would be worse than random pairs ($p < .001$). However, if observers with biases that are uncorrelated (i.e., independent pairs) or opposite (i.e., opponent pairs) with each other were paired together, their localization performance significantly outperformed the random pair performance ($p < .001$ and $p < .001$). This suggested that biases can influence the optimization of collective decisions, and for a basic localization task, pairing individuals with independent or opposite biases can both amplify the benefit of the wisdom-of-the-crowd.

It is worth concerning that Experiment 1 employed a task involving only basic and relatively low-level visual perceptual abilities. Experiment 2 and 3 aimed to further investigate whether the effect of biases in the optimization of collective decisions can be extended to more complex tasks that require higher-level perceptual and cognitive processing.

Experiment 2

Raw data for experiment 2 were retrieved from a previously published paper (Manassi et al., 2021; Experiment 1 and 2).

Method

Participants

Fifteen radiologists (11 male, 4 female, age: 27-72 years) and eleven untrained college students from University of California, Berkeley (4 male, 7 female, age: 19-21 years) participated in the experiment. As reported by Manassi and his colleagues, there were two radiologists who did not finish the experiment so their data were excluded, resulting in a total of 13 radiologists (10 male, 3 female, age: 31-72 years). Experiment procedures were approved by and conducted in accordance with the guidelines and regulations of the Institutional Review Board at University of California, Berkeley. Participants were provided with informed consent approved by the Institutional Review Board at University of California, Berkeley.

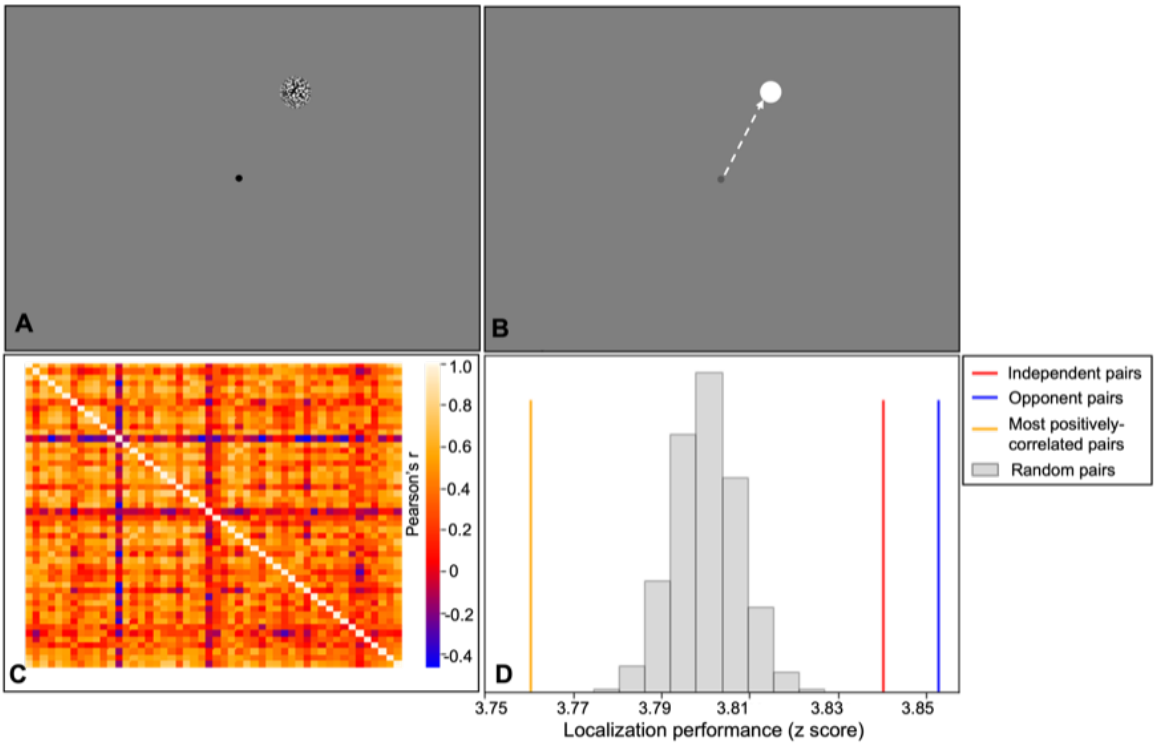

Figure 4.1: Experiment 1 paradigm and results. (a) On each trial, a noise patch enveloped within a two-dimensional Gaussian contrast aperture was presented briefly on a random location on screen. (b) After the offset of the noise patch, observers adjusted the cursor to match the location of the noise patch. (c) Pairwise correlations between all observers were calculated based on the localization errors (response – actual location). Previous studies have shown that localization errors were consistent within each observer and idiosyncratic across people, suggesting that they could be intrinsic biases of each observer. Therefore, we hypothesize that if we pair the observers who have independent (correlation closest to 0) biases, their responses should become more bias-free and accurate. Out of curiosity, we also examined the performance resulted from pairing observers who were most negatively or positively correlated. (d) Comparison between the performance of different pair observers. Results demonstrated that pairing observers that are most uncorrelated and independent is significantly better than randomly pairs ($p < .001$), suggesting that pairing the most independent observers based on their idiosyncratic biases can optimize the combined decision in visual localization. Opponent (the most negatively correlated) pairs were also significantly better than random pairs ($p < .001$), and most positively-correlated pairs would in fact extremize their shared biases and performed significantly worse than random pairs ($p < .001$).

Stimuli, Procedure and Apparatus

All stimuli were generated and presented on a 13.3 inch 2017 MacBook Pro (Apple Inc., Cupertino, CA; 1440 × 900 pixels resolution, 60 Hz refresh rate). Stimuli and the experiment were programmed using PsychoPy (Peirce, 2007, 2009). Observers viewed the stimuli binocularly at a distance of approximately 57 cm and they used the laptop keyboard to respond.

Observers were instructed to match the appearance of a light-gray random shape presented on the screen. The random shapes were generated by morphing between three anchor shapes as shown in Figure 4.2A. Between each pair of anchor shapes, 48 morph images were created, which resulted in a total of 147 random shapes in a morph continnum to use in the experiment. The shapes were approximately 3.7° width and height and they were enveloped within a two-dimensional Gaussian contrast aperture (kernel size: 1.55°).

Figure 4.2B displayed the time sequence of an example trial in the experiment. On each trial, a pseudo-randomly-selected shape superimposed on a randomly-selected mammogram image taken from The Digital Database for Screening Mammography (Heath et al., 2001; 100 mammogram candidates) was presented to observers at a random location around 4.4° eccentricity relative to a central fixation dot (0.35°). The mammogram background was set to 30% transparency. After 500 ms, the shape and the mammogram background disappeared and a noise mask composed of random Brownian noise background ($1/f^2$ spatial noise) was then presented for 1000 ms. Then, one of the shapes drawn randomly from the morph continuum appeared at the center of the screen, and observers were instructed to press the left or right arrow key and the appearance of the shape was adjusted along the morph continuum accordingly. The task for the observers is to match the shape with the previously shown random shape and they were given unlimited time to adjust their response. Once they made their decision, they pressed the spacebar to confirm their response. After a 250 ms inter-trial interval, the next trial began.

In Experiment 2, observers completed 3 blocks with 85 trials in each block. They also completed a practice block of 10 trials prior to the experiment and these practice trials were not included in the data analysis. Mean adjustment time across all observers was 3240±804 ms.

Data Analysis

Though the original paper (Manassi et al., 2021) aimed to compare the performance between radiologists and naïve observers, this was not the purpose of our current study so radiologists and naïve observers were treated as one whole group of participants.

Similar to Experiment 1, raw data of each observer were split into two halves, where one half was used to evaluate the pairwise associations between observers, and the other half served as a test set to compare the shape recognition performance of each pair. For half one, we calculated the adjustment errors as the closest distance between the targe shapes and observers' adjustment responses along the morph continuum. These residual errors

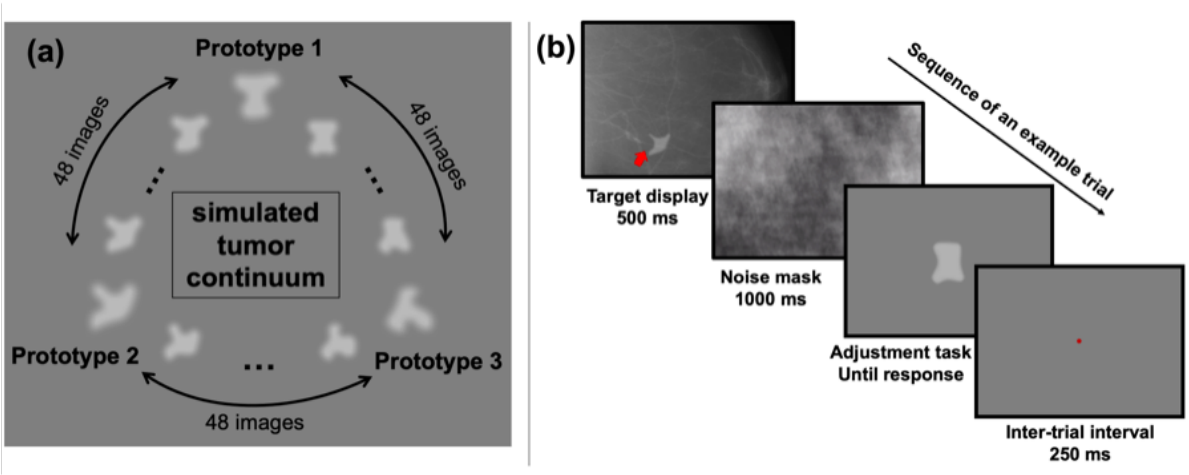

Figure 4.2: Experiment 2 stimuli, paradigm and results. Stimuli and paradigm were adapted from Manassi et al. (2021). (a) Experiment 2 used 147 random shapes generated from three anchor shapes (outline by dotted lines). (b) On each trial, a randomly chosen morph shape (pointed by the red arrow) superimposed onto a real mammograph was presented to observers for 500 ms, followed by a 1000-ms full-screen noise mask. Then one of the shapes taken randomly from the morph continuum as demonstrated in (a) appeared at the center of the screen. Observers pressed the left or right arrow key to adjust the appearance of this morph shape to match with the previously shown target and they pressed the space bar to confirm their selection. (c) We estimated the shape recognition biases from each observer and paired them in different ways (independent, opponent, most positively-correlated and random) based on the similarity of their biases. Results demonstrated that pairing observers whose biases were uncorrelated and independent from each other can outperform random pair performance ($p < .01$, permutation test using Bonferroni-corrected alpha, $\alpha_B = .0125$), while most positively-correlated or negatively-correlated pairs did not differ significantly from the random pairs ($p = .40$ and $p = .06$ respectively, $\alpha_B = .0125$).

have been shown to be serially-dependent, stimulus-specific and consistent (Manassi et al., 2021; Wang, in press), suggesting that they were observers' systematic biases in the shape recognition.

Then, pairwise Pearson's correlations were computed between these recognition biases of each pair of observers, and these 24 observers were grouped into three special pairs: 12 most positively-correlated pairs, 12 opponent pairs and 12 independent pairs, according to their correlations (see Experiment 1 Data Analysis for more details). With another half of the data, we compared the recognition performance between different ways of pairing. We took the per-stimulus responses from the two observers in a pair and averaged their responses. The averaged responses were subsequently correlated with the target morph shape IDs, and the Pearson's r value was transformed into Fisher z. Fisher z values across all 12 independent pairs were averaged to estimate the overall recognition performance of the independent pairing, same as the most positively-correlated pairs and the opponent pairs. Random pair permutation was simulated the same way as Experiment 1 for 10,000 times to estimate a permutation distribution, and we again computed the average recognition performance of single observers, which allowed us to examine whether wisdom-of-the-crowd existed in this random shape recognition task.

Results

The distribution of random pairs' recognition performance was compared to single observers' performance, revealing a significant advantage of combining two observers' responses and thus replicated the wisdom of the crowd ($p < .001$, permutation test using a Bonferroni-corrected alpha, $\alpha_B = .0125$). For the three special pairs, results were shown in Figure 4.2C. We found that the independent pairs where observers with uncorrelated biases were combined together significantly outperformed the random pairs ($p < .01, \alpha_B = .0125$), while neither the most positively-correlated pairs nor the most negatively-correlated pairs (i.e., opponent pairs) demonstrated this enhanced benefit over random pairing ($p = .40$ and $p = .06$ respectively, $\alpha_B = .0125$).

This might seem unexpected since most would assume that just as the localization task, the most negatively-correlated pairs should perform the best, with a previous study suggested that pairwise correlations can be negatively associated with performance (Corbett & Munneke, 2018). However, we believe that there could be several plausible reasons for this difference. First of all, the localization task employed in Experiment 1 is a relatively low-level task that requires majorly visual perceptual abilities, while in Experiment 2, observers need to perform a more complex shape recognition tasks that not only involve object recognition skills as well as their working memory, so their difficulty and their required visual and cognitive skills are substantially distinctive. Secondly, previous studies have suggested that human biases in object localization tend to be more consistent in comparison to biases existed in object or face recognition (Kosovicheva and Whitney, 2017; Z. Wang et al., 2020; Canas-Bajo and Whitney, 2020; Wang, in press). The variability of the recognition biases may also limit the advantage coming from pairing individuals with exact opposite biases.

Experiment 1 and 2 have both replicated the wisdom-of-the-crowd and demonstrated that if we combine the responses from observers who have independent biases together, it can provide significant benefit for optimizing collective decisions of pairs of individuals. These results may pinpoint the importance of measuring and evaluating the biases in individual's performance in many different fields or occupations, but there is a crucial gap between these tasks and the real-world practice. In many social contexts, objectively verifiable answers were not available, while both Experiment 1 and 2 and most other studies on wisdom of the crowd have employed tasks with objectively ground truth known to the researchers (for example, the exact location of each target in Experiment 1, or the exact ID of each random shape presented in Experiment 2). Therefore, Experiment 3 aimed to fill in this gap by investigating the influence of variations in human biases on collective decisions in a task with only subjective rather than objective ground truth.

Experiment 3

Raw data for experiment 3 were retrieved from a previously published paper (Chen and Whitney, 2019; Experiment 2).

Method

Participants

203 college students from University of California, Berkeley (129 females, 2 others; age range: 18 – 37 years) participated in the experiment online. For the purpose of our current study, we only analyzed the data from 101 observers: all 51 observers who rated the fully-informed videos and all 50 observers who rated the context-only videos (see below and Chen & Whitney, 2019 for more details about the experimental conditions). Experiment procedures were approved by and conducted in accordance with the guidelines and regulations of the Institutional Review Board at University of California, Berkeley. Participants were provided with informed consent approved by the Institutional Review Board at University of California, Berkeley.

Stimuli, Procedure and Apparatus

All video clips were presented via a custom website with an embedded YouTube player. Observers watched the movies with their own laptops or monitors in a non-lab environment.

Twelve natural movie stimuli (total length: 1,214 seconds) were used in this experiment. These videos were collected from YouTube and were from Hollywood movies (see Chen and Whitney, 2019 for details about how movies were selected). Observers in the fully-informed condition watched the intact movies (Figure 4.3A, upper panel) while observers in the context-only condition watched modified movies with a chosen character masked out by a Gaussian-blurred mask (Figure 4.3A, lower panel).

Observers were instructed to track the emotion of the masked-out character in each video in the context-only condition. In the fully-informed condition, observers tracked the same character without the Gaussian mask, and before they started to watch each video, they were presented with an example frame from the video with the target character's face and body visible to instruct them to track and rate the emotion of this specific character. They adjusted their cursor within a valence-arousal affect rating grid (see Figure 4.3A for a demonstration; note that the location of valence and arousal axes was randomized to be horizontal or vertical) displayed at the center of the videos to indicate the emotion of the tracked character continuously throughout each video clip. Each video clip lasted from 58 to 178 seconds and they were presented in a random order.

Data Analysis

For each observer, their cursor locations were projected onto the valence and arousal axes and transformed into continuous ratings. Then, every 100 ms of their continuous rating data were averaged together. For the purpose of our study, the below analyses only used the binned valence ratings to examine the wisdom of the crowd and the effect of biases on it.

Since each observer rated twelve videos, the valence ratings from observers in the context-only condition were split into two halves where each half contained ratings for 6 videos. Same as the analysis in Experiments 1 and 2, special pairings were determined with one half and the emotion tracking performance was evaluated with the other half.

Pairwise correlations were calculated with the first half. Similar to Experiments 1 and 2, biases were defined as residual errors in Experiment 3. To compute the residual errors, for every observer's continuous ratings for each video, we first standardized the ratings by subtracting the mean rating of this observer and dividing by the standard deviation of this observer's ratings. Subsequently, biases were calculated as the difference between each observer's ratings and the average ratings across all individuals for this video and pairwise correlations were calculated between each observer's biases. We repeated this procedure for every video and every observer with the first half of the data, and the pairwise correlations were averaged across all six videos, which yielded the overall similarity matrix between each pair of observer across the 6 different videos. Then, three special pairs (most positively-correlated, independent and opponent pairs) were then selected using maximum weight matching algorithm (Galil, 1986; see Experiment 1 Data Analysis for more details).

Before the evaluation of the emotion tracking performance of different pairs, we first need to find the "correct answer" or ground truth for the valence ratings. Since there is no objective ground truth, consensus between observers was used as a proxy of the ground truth. Consensus of each video was estimated by the average valence ratings from observers in the fully-informed condition, where observers watched the same but intact videos without the character masked out. Note that in all other previous analyses we only used the ratings from observers in the context-only condition, so the consensus was extracted from a separate and independent group of observers and a distinct and fully-informed experimental condition. This method ensured that the consensus was relatively objectively, was not entangled or

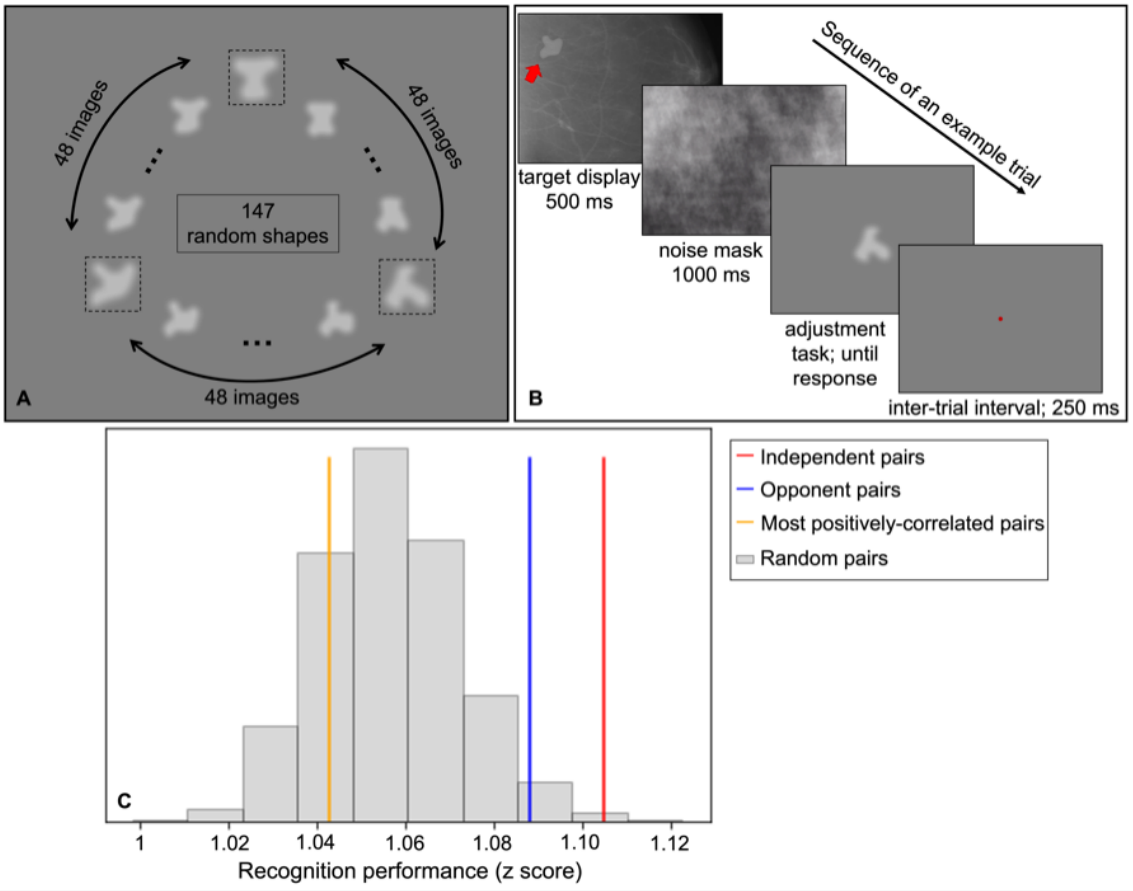

Figure 4.3: Experiment 3 stimuli, paradigm and results. Stimuli and paradigm were adapted from Chen and Whitney (2019). (a) In Experiment 3, observers watched natural movies including Hollywood movies, home videos, and documentaries. At the same time, they continuously reported the affect (valence and arousal) of a character (for example, denoted by the red dashed ellipse in Fig. 4.3a; note that the ellipse was not presented in the actual experiment) in the movie using their cursor within the response grid. The arousal and valence response grid was shown on top of the video. Each observer was presented with either fully-informed (top panel) or context-only (bottom panel) videos. Context-only videos blurred out the face and body of the chosen character so observers need to infer the character's emotion using only context information. (b) Since there was no physical ground truth defined in this emotional tracking task, we used the consensus (collective responses) from all observers in the fully-informed videos as the ground truth, and then emotion tracking performance was estimated as the similarity between the responses in the context-only videos and the consensus of fully-informed videos. Results again demonstrated the compelling advantage of pairing individuals who have independent biases from each other, which significantly outperformed random pairs ($p = .001$, permutation test using a Bonferroni-corrected alpha, $\alpha_B = .0125$), while the most positively or negatively correlated observers did not ($p = .66$ and $p = .68$, respectively; $\alpha_B = .0125$).

contaminated by the responses from observers in the context-only condition and thus could serve as a proxy for ground truth.

Afterwards, with the remaining half of the valence ratings from context-only condition, the emotion tracking performance of single observer, random pairs and three special pairs could be compared together. To estimate the overall emotion tracking performance of the opponent pairs (i.e., the average correlation across all pairs was most negative), for each pair, their continuous valence ratings for each video were averaged together and the averaged ratings of this pair were correlated with the consensus of the same video estimated by the group average ratings from the fully-informed condition. Then the yielded Pearson's r was transformed to Fisher z, which demonstrated their emotion tracking performance in this video. The Fisher z values of this pair across all six videos were subsequently averaged together. This computation was repeated on all 25 pairs (in total 50 observers in the context-only condition) and their Fisher z values were then averaged to estimate the overall emotion tracking performance of opponent pairs. Same procedures were also applied on most positively-correlated pairs and independent pairs.

To estimate the random pair performance distribution, on each iteration, observers were randomly divided into 25 pairs and their overall emotion tracking performance was calculated as described above. This procedure was repeated for 10,000 times to yield a permutation distribution of the random pairing performance. We also correlated each single observer's emotion tracking responses to the group consensus and averaged their correlations in order to test whether there could be a wisdom-of-the-crowd benefit.

Results

Wisdom-of-the-crowd was replicated in Experiment 3, which was supported by the significantly better performance of random pairs compared to single observers ($p < .001$; permutation test using a Bonferroni-corrected alpha,$\alpha_B = .0125$). This suggested that collective decisions can be beneficial for real-life situations when ground truth is not accessible or objectively defined. More importantly, results again revealed the enhanced benefit of pairing observers whose biases were uncorrelated and independent with each other compared to the random pairing technique ($p = .001, \alpha_B = .0125$). Besides, similar to Experiment 2, we found that pairing observers with negatively-correlated or positively-correlated biases could not significantly exceed the emotion tracking performance of random pairs ($p = .66$ and $p = .68$ respectively, $\alpha_B = .0125$).

As indicated by the consistent benefit of pairing independent observers across all three experiments, ranging from low-level visual localization task and higher-level shape recognition and emotion tracking tasks, we believed that combining the responses from individuals whose biases are uncorrelated and thus independent with each other is helpful for making better collective decisions. This technique can be widely applied to critical real-world applications such as TSA and radiological screening where usually a pair of readers are employed and their assessment was combined to make the final decision.

Discussion

Large quantities of critical decisions were made based on human expertise, ranging from everyday activities such as driving or sophisticated tasks including aircraft operations. However, not everyone possesses the same level of expertise, even among professionals (e.g., Elmore et al., 1994; Feldman et al., 1995; Beam et al., 1996; Elmore et al., 1998). There are two aspects that characterize the individual differences in expertise: variations in sensitivity and variations in bias. A rich body of previous research has revealed that both naïve observers and professionals have substantial individual variation in their sensitivity. For example, different radiologists can vary in their visuospatial and object recognition abilities (Smoker et al., 1984; Corry, 2011; Birchall, 2015; Langlois et al., 2015; Sunday, Donnelly, & Gauthier, 2017, 2018); naïve individuals also vary in their sensitivity when we recognize faces, despite the extensive training that we have received throughout everyday exposure (Wilmer et al., 2010). On the other hand, more recent research has also demonstrated substantial idiosyncratic variations in individual bias. For instance, in the field of vision research, these individual biases have been shown across lower-level spatial vision, mid-level motion and color perception and higher-level face recognition and medical image perception (Schütz, 2014; Wexler et al., 2015; Kosovicheva and Whitney, 2017; Kaneko et al., 2018; Emery et al., 2019; Z. Wang et al., 2020; Canas-Bajo and Whitney, 2020; Cretenoud et al., 2020; Cretenoud et al., 2021; Wang, in press). These two aspects of individual expertise may not be completely independent (Wei & Stocker, 2017), but they are undoubtedly dissociable and together characterize the envelope of human performance.

To overcome the weakness existing in individual human's expertise, past research has suggested a strategy called "wisdom of the crowd", in which having multiple observers to together form a collective decision can be a substantial benefit in situations where there is low signal, high noise, or high uncertainty. It has been shown that averaging across multiple observers improves the accuracy of judgments compared to single observers in visual search, detection, discrimination, classification, recognition, and many other tasks (e.g., Bruce, 1935; Metz and Shen, 1992; Wallsten et al., 1997; Armstrong, 2001; Surowiecki, 2004; Jiang et al., 2006; Juni and Eckstein, 2017). In some places, this multiple-observer technique is actively employed in real-world applications, for example in radiological screening in the United Kingdom. By averaging across individuals, previous studies have suggested that the benefit of this technique comes from the reduced random noises that are assumed to influence the sensitivity of individual humans, and thus yields a more accurate decision (Metz and Shen, 1992; Jiang et al., 2006; Juni and Eckstein, 2017).

However, as we have discussed earlier, human expertise and performance is not just limited by sensitivity, but also by bias, which is systematic rather than random. Though the prevalence of bias in human has been widely demonstrated (Schütz, 2014; Wexler et al., 2015; Kosovicheva and Whitney, 2017; Kaneko et al., 2018; Emery et al., 2019; Z. Wang et al., 2020; Canas-Bajo and Whitney, 2020; Cretenoud et al., 2020; Cretenoud et al., 2021; Wang, in press), bias is not included, incorporated or addressed by any current models or applications of the wisdom of the crowd. Our results directly shed light on this question, revealing

that human biases played an important role in the optimization of the wisdom-of-the-crowd. Throughout three distinct tasks ranging from low-level object localization to higher-level shape recognition and emotion tracking tasks, results constantly demonstrated that pairing independent individuals whose biases are most uncorrelated with each other can improve their combined performance compared to random pairing, even under situations where "correct answers" were subjective rather than objective such as an emotion tracking task. This clearly raised the importance to understand and measure the idiosyncratic biases that widely exist in human performance, and it provided direct implications for crucial real-world occupations such as employing pairs of readers in radiology, pathology or TSA screeners, as well as situations where subjective assessment from multiple people are combined to make a collective decision, including judging Olympic games, grading academic essays and forming a trial jury.

Unlike the consistent benefit of independent pairs, our results only showed the benefit of negatively-correlated pairs (i.e., opponent pairs) in the localization task but failed to extend it to higher-level object recognition or subjective emotion tracking task. Similarly, we only found the disadvantage of coupling the most positively correlated pairs in the localization task but not in the others. These results can be attributed to the decrease consistency in the perceptual biases in higher-level visual cognitive tasks as revealed by previous studies (Wexler et al., 2015; Kosovicheva and Whitney, 2017; Canas-Bajo and Whitney, 2020; Wang, in press). Since we used separate datasets to select the pairs and to test the paired performance, both opponent pairs and the most positively-correlated pairs did not survive this cross validation in the latter two tasks where more complex and multi-facet perceptual biases were involved. However, the real-world tasks that observers perform usually involve multi-level interactions with complex stimuli, which are undoubtedly closer to the higher-level shape recognition (Experiment 2) and emotion tracking (Experiment 3) tasks compared to the simple localization task (Experiment 1). Therefore, our findings suggested that pairing the individuals whose biases are uncorrelated with each other would be a more reliable and consistent advantage regardless of task difficulty.

Past literature has also proposed some other techniques to incorporate the variations in human responses and achieve a more accurate collective decision, such as the Cultural Consensus Theory (CCT; see Weller, 2007 for a review). We tested utilizing CCT to determine the pairing of individuals, but results demonstrated that CCT cannot significantly improve the paired performance in any of the three tasks (see Supplemental Figures S6 for details). One of the major differences between our approach and CCT is that biases are explicitly evaluated and incorporated in our approach while CCT and other studies (e.g., Corbett and Munneke, 2018) did not. Therefore, it is plausible that the advantages we found could partially arise from the treatment on individual biases.

Another key measurement in the current approach lies in the "independency". In fact, previous research has repeatedly assumed that the benefit of the wisdom-of-the-crowd exists because different observers have, to some extent, statistically independent errors in their responses (Laplace, 1812; Wallsten et al., 1997; Armstrong, 2001; Surowiecki, 2004), and some recent studies have also shown that even combining independent estimates from the same

individual can boost their performance accuracy (Vul and Pashler, 2008; Herzog & Hertwig, 2009, 2014). However, biases are not considered as part of those random noise errors that are approximately normally distributed. For example, a recent study revealed that observers are characterized with idiosyncratic biases in their perceived directions of ambiguous visual stimuli, and the temporal dynamics of these biases can be partially captured by a random walk model (Wexler et al., 2015). Our study is, to our knowledge, the first study to incorporate the variations in human biases that are systematic, dynamic and not necessarily and usually not normally-distributed, into the optimization of either group decision-making or the wisdom-of-the-crowd and it fills this critical gap that previous research has largely ignored.

Though multiple tasks across different levels of visual and cognitive processing with stimuli ranging from simple and arbitrary noise patches to natural movies were employed in our study, the real-world implications of the results can be still limited since this effect has not yet been tested within real-world settings, such as radiological diagnosis or Olympic judges. Besides, though in Experiment 2, a group of expert radiologists were recruited, we did not perform a separate analysis to compare the benefit of coupling individuals with independent biases among expert radiologists or naïve observers due to the very limited size of expert subject pool (see Experiment 2, Method). Thus, an important future follow-up of the current study is to investigate whether this effect can in reality be extended to real-world tasks and to test it explicitly on experts.

To conclude, we revealed that pairing individuals based on the biases in their responses can boost their performance, and this technique can be widely applied even to situations where ground truth is subjective rather than objective, demonstrated by the emotion tracking task. Our findings highlight the importance of measuring and understanding individual differences in human biases, and they have far-reaching implications for crucial real-world occupations such as radiologists, TSA screeners, Olympic judges and trial juries where subjective assessment from multiple people are combined to make a collective decision.

Chapter 5

Conclusion

Throughout the three chapters in this dissertation, we demonstrated that the individual differences in human perceptual biases matter. Chapter 2 revealed that idiosyncratic localization biases could be associated with the acuity and they can predict the perceived size of objects for each individual. This indicated that these spatial biases might have the same origins and they can propagate through different levels of visual processing. Chapter 3 extended the findings of idiosyncratic biases to a real-world application, showing that each radiologist is characterized with unique perceptual biases toward medical images and we suggested that these perceptual biases could be one of the reasons for the variations in their diagnostic performance. Chapter 4 further found that these perceptual biases could help us make better collective decisions. By combining the responses from individuals with uncorrelated biases (i.e., independent pairs), results showed that across three distinct tasks, these independent pairs can consistently optimize the collective performance. Together, we believed that our findings suggest that human visual perception is highly idiosyncratic, characterized with observer-specific and stimuli-specific perceptual biases. Understanding these idiosyncratic perceptual biases can be helpful for improving human perceptual performance, individually or collectively, and it can have far-reaching influence on various real-world applications such as career coaching, training and performance optimization, employing pairs of readers in radiological and TSA screenings, or constituting groups of judges in Olympic games and on the court.

Bibliography

Abrams, J., Nizam, A., & Carrasco, M. (2012). Isoeccentric locations are not equivalent: The extent of the vertical meridian asymmetry. *Vision research*, *52*(1), 70–78.

Afraz, A., Pashkam, M. V., & Cavanagh, P. (2010). Spatial heterogeneity in the perception of face and form attributes. *Current Biology*, *20*(23), 2112–2116.

Amendoeira, I., Perry, N., Broeders, M., de Wolf, C., Törnberg, S., Holland, R., von Karsa, L., et al. (2013). *European guidelines for quality assurance in breast cancer screening and diagnosis*. European Commission.

Andersen, R. A., Snyder, L. H., Bradley, D. C., & Xing, J. (1997). Multimodal representation of space in the posterior parietal cortex and its use in planning movements. *Annual review of neuroscience*, *20*, 303–330.

Appelle, S. (1972). Perception and discrimination as a function of stimulus orientation: The" oblique effect" in man and animals. *Psychological bulletin*, *78*(4), 266.

Armstrong, J. S. (2001). Judgmental bootstrapping: Inferring experts' rules for forecasting. In *Principles of forecasting* (pp. 171–192). Springer.

Atkinson, J., Braddick, O., & French, J. (1980). Infant astigmatism: Its disappearance with age. *Vision research*, *20*(11), 891–893.

Australia, B. (2002). National accreditation standards. breastscreen quality improvement program.

Bass, J. C., & Chiles, C. (1990). Visual skill correlation with detection of solitary pulmonary nodules. *Investigative radiology*, *25*(9), 994–997.

Beam, C. A., Layde, P. M., & Sullivan, D. C. (1996). Variability in the interpretation of screening mammograms by us radiologists: Findings from a national sample. *Archives of internal medicine*, *156*(2), 209–213.

Berg, W. A., D'Orsi, C. J., Jackson, V. P., Bassett, L. W., Beam, C. A., Lewis, R. S., & Crewson, P. E. (2002). Does training in the breast imaging reporting and data system (bi-rads) improve biopsy recommendations or feature analysis agreement with experienced breast imagers at mammography? *Radiology*, *224*(3), 871–880.

Birchall, D. (2015). Spatial ability in radiologists: A necessary prerequisite? *The British journal of radiology*, *88*(1049), 20140511.

Bobak, A. K., Hancock, P. J., & Bate, S. (2016). Super-recognisers in action: Evidence from face-matching and face memory tasks. *Applied Cognitive Psychology*, *30*(1), 81–91.

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

Brennan, P. C., Ganesan, A., Eckstein, M. P., Ekpo, E. U., Tapia, K., Mello-Thoms, C., Lewis, S., & Juni, M. Z. (2019). Benefits of independent double reading in digital mammography: A theoretical evaluation of all possible pairing methodologies. *Academic radiology*, *26*(6), 717–723.

Bruce, R. S. (1935). Group judgments in the fields of lifted weights and visual discrimination. *The Journal of Psychology*, *1*(1), 117–121.

Burr, D. C., Morrone, M. C., & Ross, J. (2001). Separate visual representations for perception and action revealed by saccadic eye movements. *Current Biology*, *11*(10), 798–802.

Canas-Bajo, T., & Whitney, D. (2020). Stimulus-specific individual differences in holistic perception of mooney faces. *Frontiers in Psychology*, *11*, 585921.

Chen, Z., & Whitney, D. (2019). Tracking the affective state of unseen persons. *Proceedings of the National Academy of Sciences*, *116*(15), 7559–7564.

Chua, K.-W., & Gauthier, I. (2020). Domain-specific experience determines individual differences in holistic processing. *Journal of Experimental Psychology: General*, *149*(1), 31.

Cohen, L., Gray, F., Meyrignac, C., Dehaene, S., & Degos, J.-D. (1994). Selective deficit of visual size perception: Two cases of hemimicropsia. *Journal of Neurology, Neurosurgery & Psychiatry*, *57*(1), 73–78.

Corbett, J. E., & Munneke, J. (2018). "it's not a tumor": A framework for capitalizing on individual diversity to boost target detection. *Psychological Science*, *29*(10), 1692–1705.

Corry, C. (2011). The future of recruitment and selection in radiology. is there a role for assessment of basic visuospatial skills? *Clinical radiology*, *66*(5), 481–483.

Cretenoud, A. F., Grzeczkowski, L., Bertamini, M., & Herzog, M. H. (2020). Individual differences in the müller-lyer and ponzo illusions are stable across different contexts. *Journal of Vision*, *20*(6), 4–4.

Cretenoud, A. F., Grzeczkowski, L., Kunchulia, M., & Herzog, M. H. (2021). Individual differences in the perception of visual illusions are stable across eyes, time, and measurement methods. *Journal of vision*, *21*(5), 26–26.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *psychometrika*, *16*(3), 297–334.

Donald, J. J., & Barnard, S. A. (2012). Common patterns in 558 diagnostic radiology errors. *Journal of medical imaging and radiation oncology*, *56*(2), 173–178.

Donovan, T., & Litchfield, D. (2013). Looking for cancer: Expertise related differences in searching and decision making. *Applied Cognitive Psychology*, *27*(1), 43–49.

Donovan, T., Litchfield, D., & Crawford, T. J. (2017). Medical image perception: How much do we understand it? *Frontiers in Psychology*, *8*, 2072.

Dotsch, R., & Todorov, A. (2012). Reverse correlating social face perception. *Social Psychological and Personality Science*, *3*(5), 562–571.

Drew, T., Vo, M. L.-H., Olwal, A., Jacobson, F., Seltzer, S. E., & Wolfe, J. M. (2013). Scanners and drillers: Characterizing expert visual search through volumetric images. *Journal of vision*, *13*(10), 3–3.

Duchaine, B., & Nakayama, K. (2006). The cambridge face memory test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia*, *44*(4), 576–585.

Duncan, R. O., & Boynton, G. M. (2003). Cortical magnification within human primary visual cortex correlates with acuity thresholds. *Neuron*, *38*(4), 659–671.

Dwass, M. (1957). Modified randomization tests for nonparametric hypotheses. *The Annals of Mathematical Statistics*, 181–187.

Edgington, E., & Onghena, P. (2007). *Randomization tests*. Chapman; Hall/CRC.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Elmore, J. G., Jackson, S. L., Abraham, L., Miglioretti, D. L., Carney, P. A., Geller, B. M., Yankaskas, B. C., Kerlikowske, K., Onega, T., Rosenberg, R. D., et al. (2009). Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy. *Radiology*, *253*(3), 641.

Elmore, J. G., Miglioretti, D. L., Reisch, L. M., Barton, M. B., Kreuter, W., Christiansen, C. L., & Fletcher, S. W. (2002). Screening mammograms by community radiologists: Variability in false-positive rates. *Journal of the National Cancer Institute*, *94*(18), 1373–1380.

Elmore, J. G., Wells, C. K., & Howard, D. H. (1998). Does diagnostic accuracy in mammography depend on radiologists' experience? *Journal of Women's Health*, *7*(4), 443–449.

Elmore, J. G., Wells, C. K., Lee, C. H., Howard, D. H., & Feinstein, A. R. (1994). Variability in radiologists' interpretations of mammograms. *New England Journal of Medicine*, *331*(22), 1493–1499.

Emery, K., Volbrecht, V., Peterzell, D., & Webster, M. (2019). Color vs. motion: Decoding perceptual representations from individual differences. *Journal of Vision*, *19*(8), 8–8.

Evans, K. K., Culpan, A.-M., & Wolfe, J. M. (2019). Detecting the "gist" of breast cancer in mammograms three years before localized signs of cancer are visible. *The British journal of radiology*, *92*(1099), 20190136.

Farah, M. J., Wilson, K. D., Drain, M., & Tanaka, J. N. (1998). What is" special" about face perception? *Psychological review*, *105*(3), 482.

Feldman, J., Smith, R., Giusti, R., DeBuono, B., Fulton, J. P., & Scott, H. D. (1995). Peer review of mammography interpretations in a breast cancer screening program. *American journal of public health*, *85*(6), 837–839.

Findlay, L. C., & Ste-Marie, D. M. (2004). A reputation bias in figure skating judging. *Journal of Sport and Exercise Psychology*, *26*(1), 154–166.

Fischer, J., Spotswood, N., & Whitney, D. (2011). The emergence of perceived position in the visual system. *Journal of cognitive neuroscience*, *23*(1), 119–136.

Fisher, R. A. (1992). Statistical methods for research workers. In *Breakthroughs in statistics* (pp. 66–70). Springer.

Fletcher, J. G., Chen, M.-H., Herman, B. A., Johnson, C. D., Toledano, A., Dachman, A. H., Hara, A. K., Fidler, J. L., Menias, C. O., Coakley, K. J., et al. (2010). Can radiologist training and testing ensure high performance in ct colonography? lessons from the national ct colonography trial. *AJR. American journal of roentgenology*, *195*(1), 117.

Galil, Z. (1986). Efficient algorithms for finding maximum matching in graphs. *ACM Computing Surveys (CSUR)*, *18*(1), 23–38.

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*(2), 191–197.

Germine, L., Russell, R., Bronstad, P. M., Blokland, G. A., Smoller, J. W., Kwok, H., Anthony, S. E., Nakayama, K., Rhodes, G., & Wilmer, J. B. (2015). Individual aesthetic preferences for faces are shaped mostly by environments, not genes. *Current Biology*, *25*(20), 2684–2689.

Gescheider, G. A. (2013). *Psychophysics: The fundamentals*. Psychology Press.

Goh, J. X., Hall, J. A., & Rosenthal, R. (2016). Mini meta-analysis of your own studies: Some arguments on why and a primer on how. *Social and Personality Psychology Compass*, *10*(10), 535–549.

Gomez, J., Pestilli, F., Witthoft, N., Golarai, G., Liberman, A., Poltoratski, S., Yoon, J., & Grill-Spector, K. (2015). Functionally defined white matter reveals segregated pathways in human ventral temporal cortex associated with category-specific processing. *Neuron*, *85*(1), 216–227.

Greenwood, J. A., Szinte, M., Sayim, B., & Cavanagh, P. (2017). Variations in crowding, saccadic precision, and spatial localization reveal the shared topology of spatial vision. *Proceedings of the National Academy of Sciences*, *114*(17), E3573–E3582.

Griffin, I. C., & Nobre, A. C. (2003). Orienting attention to locations in internal representations. *Journal of cognitive neuroscience*, *15*(8), 1176–1194.

Grzeczkowski, L., Clarke, A. M., Francis, G., Mast, F. W., & Herzog, M. H. (2017). About individual differences in vision. *Vision research*, *141*, 282–292.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*(5539), 2425–2430.

Heath, M., Bowyer, K., Kopans, D., Moore, R., & Jr, P. K. (2001). *The digital database for screening mammography*. Digital Mammography.

Herman, P. G., & Hessel, S. J. (1975). Accuracy and its relationship to experience in the interpretation of chest radiographs. *Investigative Radiology*, *10*(1), 62–67.

Herzog, S. M., & Hertwig, R. (2009). The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychological Science*, *20*(2), 231–237.

Herzog, S. M., & Hertwig, R. (2014). Harnessing the wisdom of the inner crowd. *Trends in cognitive sciences*, *18*(10), 504–506.

Himmelberg, M. M., Winawer, J., & Carrasco, M. (2022). Linking individual differences in human primary visual cortex to contrast sensitivity around the visual field. *bioRxiv*, 2021–10.

Imhoff, R., Dotsch, R., Bianchi, M., Banse, R., & Wigboldus, D. H. (2011). Facing europe: Visualizing spontaneous in-group projection. *Psychological science*, *22*(12), 1583–1590.

International Skating Union. (2022). Special regulations & technical rules: Single & pair skating and ice dance 2022. https://www.isu.org/figure-skating/rules/fsk-regulations-rules/file

Irwin, D. E. (1991). Information integration across saccadic eye movements. *Cognitive psychology*, *23*(3), 420–456.

Jack, R. E., Caldara, R., & Schyns, P. G. (2012). Internal representations reveal cultural diversity in expectations of facial expressions of emotion. *Journal of Experimental Psychology: General*, *141*(1), 19.

Jiang, Y., Metz, C. E., Nishikawa, R. M., & Schmidt, R. A. (2006). Comparison of independent double readings and computer-aided diagnosis (cad) for the diagnosis of breast calcifications. *Academic radiology*, *13*(1), 84–94.

Juni, M. Z., & Eckstein, M. P. (2017). The wisdom of crowds for visual search. *Proceedings of the National Academy of Sciences*, *114*(21), E4306–E4315.

Kanai, R., & Rees, G. (2011). The structural basis of inter-individual differences in human behaviour and cognition. *Nature Reviews Neuroscience*, *12*(4), 231–242.

Kaneko, S., Murakami, I., Kuriki, I., & Peterzell, D. H. (2018). Individual variability in simultaneous contrast for color and brightness: Small sample factor analyses reveal separate induction processes for short and long flashes. *i-Perception*, *9*(5), 2041669518800507.

Kansagra, A. P., Liu, K., & John-Paul, J. Y. (2016). Disruption of radiologist workflow. *Current Problems in Diagnostic Radiology*, *45*(2), 101–106.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of neuroscience*, *17*(11), 4302–4311.

Klein, S. B., Gabriel, R. H., Gangi, C. E., & Robertson, T. E. (2008). Reflections on the self: A case study of a prosopagnosic patient. *Social Cognition*, *26*(6), 766–777.

Kosovicheva, A., & Whitney, D. (2017). Stable individual signatures in object localization. *Current Biology*, *27*(14), R700–R701.

Krupinski, E. A. (1996). Visual scanning patterns of radiologists searching mammograms. *Academic radiology*, *3*(2), 137–144.

Krupinski, E. A. (2010). Current perspectives in medical image perception. *Attention, Perception, & Psychophysics*, *72*(5), 1205–1217.

Kundel, H. L. (1989). Perception errors in chest radiography. *Seminars in Respiratory Medicine*, *10*(03), 203–210.

Kundel, H. L. (2006). History of research in medical image perception. *Journal of the American college of radiology*, *3*(6), 402–408.

Kundel, H. L., & La Follette Jr, P. S. (1972). Visual search patterns and experience with radiological images. *Radiology*, *103*(3), 523–528.

Kundel, H. L., Nodine, C. F., & Carmody, D. (1978). Visual scanning, pattern recognition and decision-making in pulmonary nodule detection. *Investigative radiology*, *13*(3), 175–181.

Langlois, J., Wells, G. A., Lecourtois, M., Bergeron, G., Yetisir, E., & Martin, M. (2015). Spatial abilities of medical graduates and choice of residency programs. *Anatomical Sciences Education*, *8*(2), 111–119.

Laplace, P. S. (1812). Marquis de. *A Philosophical Essay on Probabilities*.

Lazarus, E., Mainiero, M. B., Schepps, B., Koelliker, S. L., & Livingston, L. S. (2006). Bi-rads lexicon for us and mammography: Interobserver variability and positive predictive value. *Radiology*, *239*(2), 385–391.

Lazic, S. E. (2010). The problem of pseudoreplication in neuroscientific studies: Is it affecting your analysis? *BMC neuroscience*, *11*(1), 1–17.

Levi, D. M., & Klein, S. A. (1985). Vernier acuity, crowding and amblyopia. *Vision research*, *25*(7), 979–991.

Linacre, J. M. (2009). Local independence and residual covariance: A study of olympic figure skating ratings. *Journal of applied measurement*, *10*(2), 157–169.

Linver, M., Paster, S., Rosenberg, R., Key, C., Stidley, C., & King, W. (1992). Improvement in mammography interpretation skills in a community radiology practice after dedicated teaching courses: 2-year medical audit of 38,633 cases. *Radiology*, *184*(1), 39–43.

Low, F. N. (1943). The peripheral visual acuity of 100 subjects. *American Journal of Physiology-Legacy Content*, *140*(1), 83–88.

Manassi, M., Ghirardo, C., Canas-Bajo, T., Ren, Z., Prinzmetal, W., & Whitney, D. (2021). Serial dependence in the perceptual judgments of radiologists. *Cognitive research: principles and implications*, *6*(1), 1–13.

Manassi, M., Kristjánsson, Á., & Whitney, D. (2019). Serial dependence in a simulated clinical visual search task. *Scientific reports*, *9*(1), 1–10.

Manly, B. F. (2018). *Randomization, bootstrap and monte carlo methods in biology: Texts in statistical science*. chapman; hall/CRC.

Manning, D., Ethell, S., Donovan, T., & Crawford, T. (2006). How do radiologists do it? the influence of experience and training on searching for chest nodules. *Radiography*, *12*(2), 134–142.

Maurer, D., Le Grand, R., & Mondloch, C. J. (2002). The many faces of configural processing. *Trends in cognitive sciences*, *6*(6), 255–260.

Maus, G. W., Ward, J., Nijhawan, R., & Whitney, D. (2013). The perceived position of moving objects: Transcranial magnetic stimulation of area mt+ reduces the flash-lag effect. *Cerebral cortex*, *23*(1), 241–247.

McGraw, P. V., Walsh, V., & Barrett, B. T. (2004). Motion-sensitive neurones in v5/mt modulate perceived spatial position. *Current Biology*, *14*(12), 1090–1093.

McKee, S. P., Silverman, G. H., & Nakayama, K. (1986). Precise velocity discrimination despite random variations in temporal frequency and contrast. *Vision research*, *26*(4), 609–619.

Mercan, E., Shapiro, L. G., Brunyé, T. T., Weaver, D. L., & Elmore, J. G. (2018). Characterizing diagnostic search patterns in digital breast pathology: Scanners and drillers. *Journal of digital imaging*, *31*(1), 32–41.

Metz, C. E., & Shen, J.-H. (1992). Gains in accuracy from replicated readings of diagnostic images: Prediction and assessment in terms of roc analysis. *Medical Decision Making*, *12*(1), 60–75.

Mishkin, M., Ungerleider, L. G., & Macko, K. A. (1983). Object vision and spatial vision: Two cortical pathways. *Trends in neurosciences*, *6*, 414–417.

Mohindra, I., Held, R., Gwiazda, J., & Brill, S. (1978). Astigmatism in infants. *Science*, *202*(4365), 329–331.

Molins, E., Macià, F., Ferrer, F., Maristany, M.-T., & Castells, X. (2008). Association between radiologists' experience and accuracy in interpreting screening mammograms. *BMC health services research*, *8*(1), 1–10.

Mollon, J. D., Bosten, J. M., Peterzell, D. H., & Webster, M. A. (2017). Individual differences in visual science: What can be learned and what is good experimental practice? *Vision Research*, *141*, 4–15.

Moon, K., Kim, S., Kim, J., Kim, H., & Ko, Y.-g. (2020). The mirror of mind: Visualizing mental representations of self through reverse correlation. *Frontiers in Psychology*, *11*, 1149.

Morgan, M., & Regan, D. (1987). Opponent model for line interval discrimination: Interval and vernier performance compared. *Vision Research*, *27*(1), 107–118.

Morgan, M., Watamaniuk, S., & McKee, S. (2000). The use of an implicit standard for measuring discrimination thresholds. *Vision research*, *40*(17), 2341–2349.

Moscovitch, M., Winocur, G., & Behrmann, M. (1997). What is special about face recognition? nineteen experiments on a person with visual object agnosia and dyslexia but normal face recognition. *Journal of cognitive neuroscience*, *9*(5), 555–604.

Moutsiana, C., De Haas, B., Papageorgiou, A., Van Dijk, J. A., Balraj, A., Greenwood, J. A., & Schwarzkopf, D. S. (2016). Cortical idiosyncrasies predict the perception of object size. *Nature communications*, *7*(1), 1–12.

Olivers, C. N., Peters, J., Houtkamp, R., & Roelfsema, P. R. (2011). Different states in visual working memory: When it guides attention and when it does not. *Trends in cognitive sciences*, *15*(7), 327–334.

Peirce, J. W. (2007). Psychopy—psychophysics software in python. *Journal of neuroscience methods*, *162*(1-2), 8–13.

Peirce, J. W. (2009). Generating stimuli for neuroscience using psychopy. *Frontiers in neuroinformatics*, *2*, 10.

Pelli, D. G., & Farell, B. (1995). Psychophysical methods. *Handbook of optics*, *1*, 29–1.

Peterzell, D. H., Werner, J. S., & Kaplan, P. S. (1995). Individual differences in contrast sensitivity functions: Longitudinal study of 4-, 6-and 8-month-old human infants. *Vision research*, *35*(7), 961–979.

Pickersgill, N. A., Vetter, J. M., Raval, N. S., Andriole, G. L., Shetty, A. S., Ippolito, J. E., & Kim, E. H. (2019). The accuracy of prostate magnetic resonance imaging interpretation: Impact of the individual radiologist and clinical factors. *Urology*, *127*, 68–73.

Pointer, J., & Hess, R. (1989). The contrast sensitivity gradient across the human visual field: With emphasis on the low spatial frequency range. *Vision research*, *29*(9), 1133–1151.

Previc, F. H. (1990). Functional specialization in the lower and upper visual fields in humans: Its ecological origins and neurophysiological implications. *Behavioral and Brain Sciences*, *13*(3), 519–542.

Prinzmetal, W., Amiri, H., Allen, K., & Edwards, T. (1998). Phenomenology of attention: I. color, location, orientation, and spatial frequency. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(1), 261.

Quekel, L. G., Kessels, A. G., Goei, R., & van Engelshoven, J. M. (1999). Miss rate of lung cancer on the chest radiograph in clinical practice. *Chest*, *115*(3), 720–724.

Ren, Z., Yu, S. X., & Whitney, D. (2021). Controllable medical image generation via generative adversarial networks. *Electronic Imaging*, *2021*(11), 112–1.

Richler, J. J., Tomarken, A. J., Sunday, M. A., Vickery, T. J., Ryan, K. F., Floyd, R. J., Sheinberg, D., Wong, A. C.-N., & Gauthier, I. (2019). Individual differences in object recognition. *Psychological Review*, *126*(2), 226.

Rosen, T., Bloemen, E. M., Harpe, J., Sanchez, A. M., Mennitt, K. W., McCarthy, T. J., Nicola, R., Murphy, K., LoFaso, V. M., Flomenbaum, N., et al. (2016). Radiologists' training, experience, and attitudes about elder abuse detection. *AJR. American journal of roentgenology*, *207*(6), 1210.

Rosenthal, R. (1978). Combining results of independent studies. *Psychological bulletin*, *85*(1), 185.

Rossion, B. (2013). The composite face illusion: A whole window into our understanding of holistic face perception. *Visual Cognition*, *21*(2), 139–253.

Russell, R., Chatterjee, G., & Nakayama, K. (2012). Developmental prosopagnosia and super-recognition: No special role for surface reflectance processing. *Neuropsychologia*, *50*(2), 334–340.

Russell, R., Duchaine, B., & Nakayama, K. (2009). Super-recognizers: People with extraordinary face recognition ability. *Psychonomic bulletin & review*, *16*(2), 252–257.

Samei, E., & Krupinski, E. (2009). The handbook of medical image perception and techniques. *The Handbook of Medical Image Perception and Techniques*.

Samei, E., & Krupinski, E. A. (2018). *The handbook of medical image perception and techniques*. Cambridge University Press.

Schütz, A. C. (2014). Interindividual differences in preferred directions of perceptual and motor decisions. *Journal of vision*, *14*(12), 16–16.

Sha, L. Z., Toh, Y. N., Remington, R. W., & Jiang, Y. V. (2020). Perceptual learning in the identification of lung cancer in chest radiographs. *Cognitive Research: Principles and Implications*, *5*(1), 1–13.

Shakeshaft, N. G., & Plomin, R. (2015). Genetic specificity of face recognition. *Proceedings of the National Academy of Sciences*, *112*(41), 12887–12892.

Smoker, W., Berbaum, K., Luebke, N., & Jacoby, C. (1984). Spatial perception testing in diagnostic radiology. *American journal of roentgenology*, *143*(5), 1105–1109.

Song, C., Schwarzkopf, D. S., Kanai, R., & Rees, G. (2015). Neural population tuning links visual cortical anatomy to human visual perception. *Neuron*, *85*(3), 641–656.

Song, S., Garrido, L., Nagy, Z., Mohammadi, S., Steel, A., Driver, J., Dolan, R. J., Duchaine, B., & Furl, N. (2015). Local but not long-range microstructural differences of the ventral temporal cortex in developmental prosopagnosia. *Neuropsychologia*, *78*, 195–206.

Sonn, G. A., Fan, R. E., Ghanouni, P., Wang, N. N., Brooks, J. D., Loening, A. M., Daniel, B. L., To'o, K. J., Thong, A. E., & Leppert, J. T. (2019). Prostate magnetic resonance imaging interpretation varies substantially across radiologists. *European urology focus*, *5*(4), 592–599.

Soto, F. A. (2019). Categorization training changes the visual representation of face identity. *Attention, Perception, & Psychophysics*, *81*(5), 1220–1227.

Stansbury, D. E., Naselaris, T., & Gallant, J. L. (2013). Natural scene statistics account for the representation of scene categories in human visual cortex. *Neuron*, *79*(5), 1025–1034.

Ste-Marie, D. M., & Lee, T. D. (1991). Prior processing effects on gymnastic judging. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(1), 126.

Stevens, S. S. (1958). Problems and methods of psychophysics. *Psychological bulletin*, *55*(4), 177.

Sunday, M. A., Donnelly, E., & Gauthier, I. (2017). Individual differences in perceptual abilities in medical imaging: The vanderbilt chest radiograph test. *Cognitive Research: Principles and Implications*, *2*(1), 1–10.

Sunday, M. A., Donnelly, E., & Gauthier, I. (2018). Both fluid intelligence and visual object recognition ability relate to nodule detection in chest radiographs. *Applied Cognitive Psychology*, *32*(6), 755–762.

Surowiecki, J. (2004). The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business. *Economies, Societies and Nations*, *296*(5).

Sutherland, C. A., Burton, N. S., Wilmer, J. B., Blokland, G. A., Germine, L., Palermo, R., Collova, J. R., & Rhodes, G. (2020). Individual differences in trust evaluations are shaped mostly by environments, not genes. *Proceedings of the National Academy of Sciences*, *117*(19), 10218–10224.

Suzuki, S., & Cavanagh, P. (1997). Focused attention distorts visual space: An attentional repulsion effect. *Journal of Experimental Psychology: Human Perception and Performance*, *23*(2), 443.

Tan, A., Freeman, D. H., Goodwin, J. S., & Freeman, J. L. (2006). Variation in false-positive rates of mammography reading among 1067 radiologists: A population-based assessment. *Breast cancer research and treatment*, *100*(3), 309–318.

Taylor-Phillips, S., & Stinton, C. (2020). Double reading in breast cancer screening: Considerations for policy-making. *The British Journal of Radiology*, *93*(1106), 20190610.

Team, R. C. (2013). R: A language and environment for statistical computing.

Theodoropoulos, J. S., Andreisek, G., Harvey, E. J., & Wolin, P. (2010). Magnetic resonance imaging and magnetic resonance arthrography of the shoulder: Dependence on

the level of training of the performing radiologist for diagnostic accuracy. *Skeletal radiology*, *39*(7), 661–667.

Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. *Network: computation in neural systems*, *14*(3), 391.

Van Such, M., Lohr, R., Beckman, T., & Naessens, J. M. (2017). Extent of diagnostic agreement among medical referrals. *Journal of evaluation in clinical practice*, *23*(4), 870–874.

Van Tubergen, A., Heuft-Dorenbosch, L., Schulpen, G., Landewé, R., Wijers, R., Van Der Heijde, D., van Engelshoven, J., & van der Linden, S. (2003). Radiographic assessment of sacroiliitis by radiologists and rheumatologists: Does training improve quality? *Annals of the rheumatic diseases*, *62*(6), 519–525.

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, *19*(7), 645–647.

Waite, S., Grigorian, A., Alexander, R. G., Macknik, S. L., Carrasco, M., Heeger, D. J., & Martinez-Conde, S. (2019). Analysis of perceptual expertise in radiology–current knowledge and a new perspective. *Frontiers in human neuroscience*, *13*, 213.

Wallsten, T. S., Budescu, D. V., Erev, I., & Diederich, A. (1997). Evaluating and combining subjective probability estimates. *Journal of Behavioral Decision Making*, *10*(3), 243–268.

Wang. (in press). Idiosyncratic biases in the perception of medical images. *Frontiers in Psychology*.

Wang, H., & Levi, D. M. (1994). Spatial integration in position acuity. *Vision Research*, *34*(21), 2859–2877.

Wang, R., Li, J., Fang, H., Tian, M., & Liu, J. (2012). Individual differences in holistic processing predict face recognition ability. *Psychological science*, *23*(2), 169–177.

Wang, Z., Murai, Y., & Whitney, D. (2020). Idiosyncratic perception: A link between acuity, perceived position and apparent size. *Proceedings of the Royal Society B*, *287*(1930), 20200825.

Wei, X.-X., & Stocker, A. A. (2017). Lawful relation between perceptual bias and discriminability. *Proceedings of the National Academy of Sciences*, *114*(38), 10244–10249.

Weller, S. C. (2007). Cultural consensus theory: Applications and frequently asked questions. *Field methods*, *19*(4), 339–368.

Westheimer, G. (1975). Visual acuity and hyperacuity. *Investigative ophthalmology*, *14*(8), 570–572.

Wexler, M., Duyck, M., & Mamassian, P. (2015). Persistent states in vision break universality and time invariance. *Proceedings of the National Academy of Sciences*, *112*(48), 14990–14995.

Whissell, R., Lyons, S., Wilkinson, D., & Whissell, C. (1993). National bias in judgments of olympic-level skating. *Perceptual and motor skills*, *77*(2), 355–358.

Whitaker, D., & MacVEIGH, D. (1992). Sequential mapping of weighting functions for visual location. *Spatial vision*, *6*(2), 117–131.

Whitney, D., & Cavanagh, P. (2000). Motion distorts visual space: Shifting the perceived position of remote stationary objects. *Nature neuroscience*, *3*(9), 954–959.

Wilmer, J. B. (2017). Individual differences in face recognition: A decade of discovery. *Current Directions in Psychological Science*, *26*(3), 225–230.

Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Williams, M., Loken, E., Nakayama, K., & Duchaine, B. (2010). Human face recognition ability is specific and highly heritable. *Proceedings of the National Academy of sciences*, *107*(11), 5238–5241.

Wilson, R., & Liston, J. (2011). Quality assurance guidelines for breast cancer screening radiology: Nhs breast screening programme publication number 59. *Sheffield, England: NHS Cancer Screening Programmes*.

Wolfe, J. M. (2022). How one block of trials influences the next: Persistent effects of disease prevalence and feedback on decisions about images of skin lesions in a large online study. *Cognitive Research: Principles and Implications*, *7*(1), 1–13.

Yan, K., Wang, X., Lu, L., & Summers, R. M. (2018). Deeplesion: Automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of medical imaging*, *5*(3), 036501.

Yankaskas, B. C., Klabunde, C. N., Ancelle-Park, R., Rennert, G., Wang, H., Fracheboud, J., Pou, G., Bulliard, J.-L., & Network, I. B. C. S. (2004). International comparison of performance measures for screening mammography: Can it be done? *Journal of medical screening*, *11*(4), 187–193.

Zhu, Q., Song, Y., Hu, S., Li, X., Tian, M., Zhen, Z., Dong, Q., Kanwisher, N., & Liu, J. (2010). Heritability of the specific cognitive ability of face perception. *Current Biology*, *20*(2), 137–142.

Zhu, Z., Chen, B., Na, R., Fang, W., Zhang, W., Zhou, Q., Zhou, S., Lei, H., Huang, A., Chen, T., et al. (2021). A genome-wide association study reveals a substantial genetic basis underlying the ebbinghaus illusion. *Journal of Human Genetics*, *66*(3), 261–271.

Zitzewitz, E. (2006). Nationalism in winter sports judging and its lessons for organizational decision making. *Journal of Economics & Management Strategy*, *15*(1), 67–99.

Appendix A: Supplemental Materials

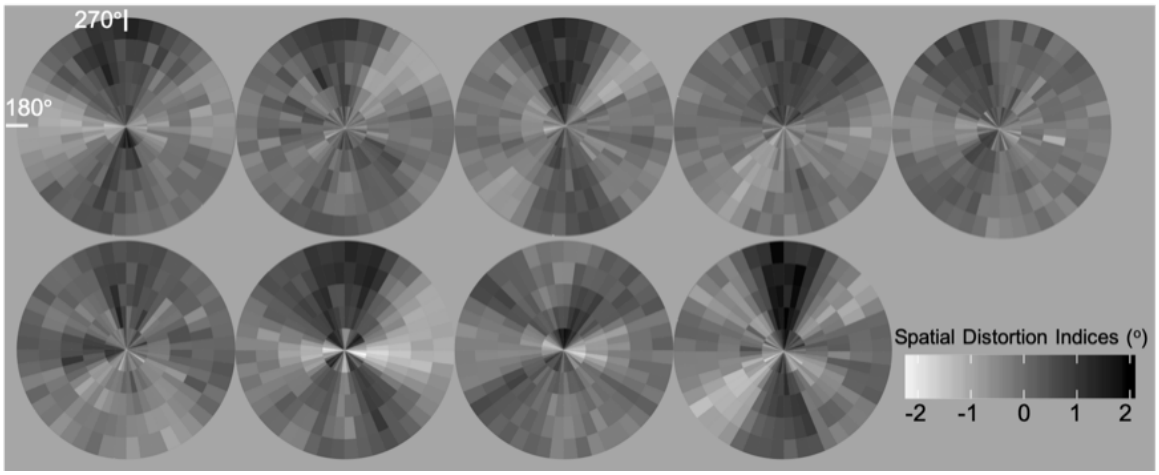

Figure S1: The gray-scale version of the spatial distortion maps reported in Chapter 2, Experiment 1. Brighter color (negative spatial distortion indices) indicates contraction of visual space and darker color (positive spatial distortion indices) represents expanded visual space.

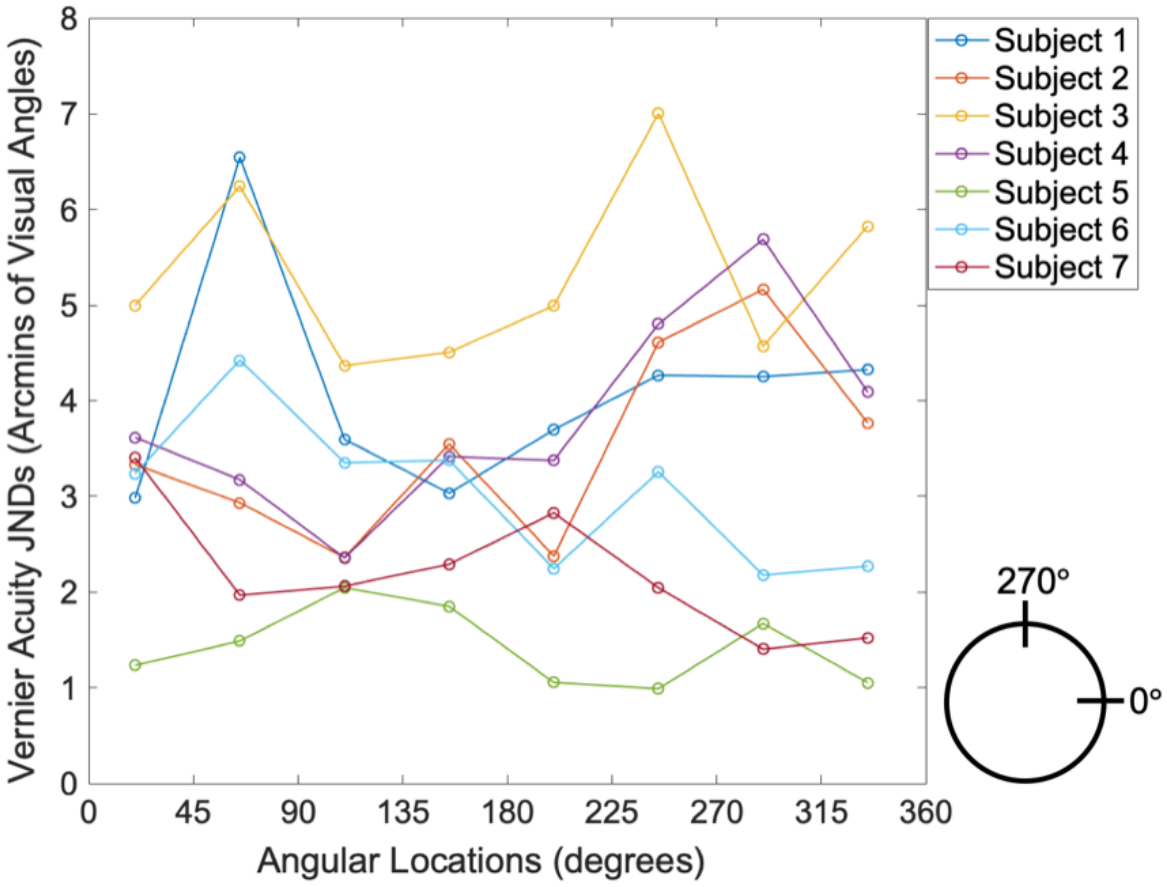

Figure S2: Change of Vernier acuity as a function of the angular locations tested for every observer. Subject 1 and Subject 4 are authors. The layout of the angular locations is shown on the bottom right corner.

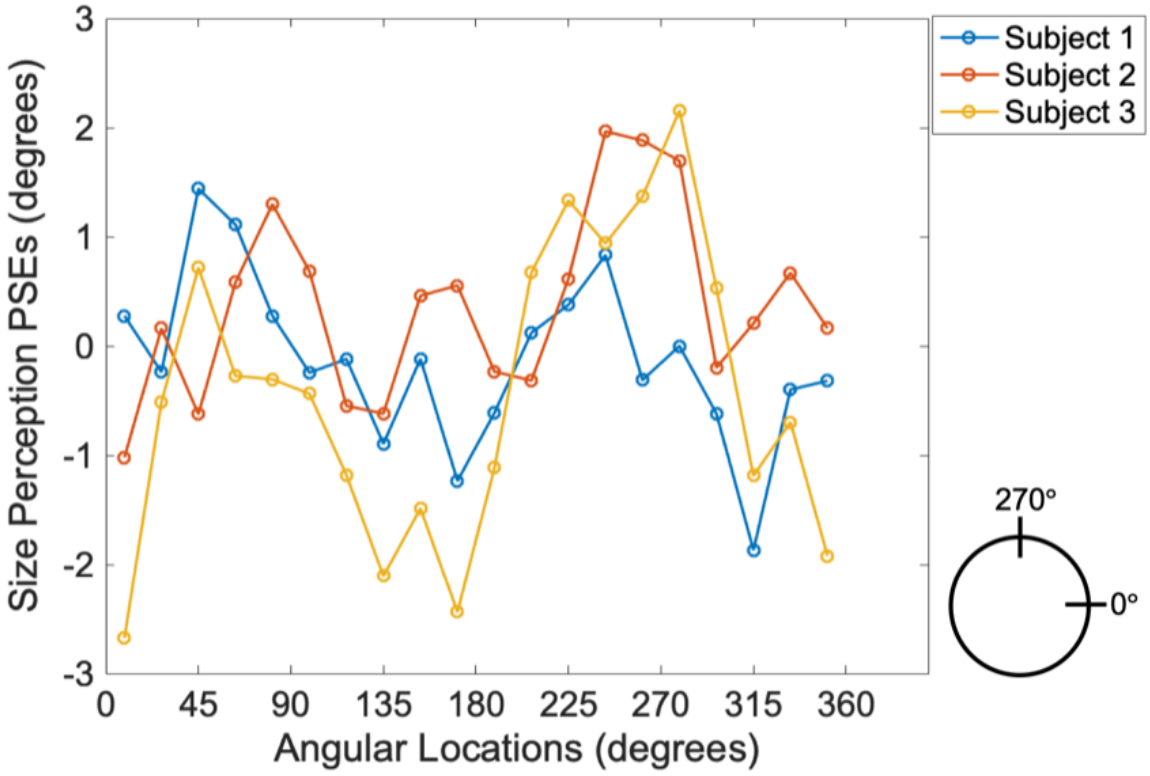

Figure S3: Correlation between spatial distortion indices and Vernier acuity JNDs for each observer. Each observer had 8 pairs of data, corresponding to 8 angular locations tested in Chapter 2, Experiment 2. Different symbols represent different observers. Lines are regression lines fitted based on each observer's data. The Pearson's correlations for individual subjects were 0.31, 0.73, 0.34, 0.55, -0.37, 0.26, 0.38 (listed in the same order as the figure legend) and the mean correlation calculated from Fisher transformation was 0.34. Note that the only observer who did not show the same trend (displayed as gray diamond) had the smallest JNDs (i.e., best acuity), so we speculated that it might be subject to a ceiling effect. This could affect the measured variability of Vernier acuity across different locations and thus influence the correlation calculated based on it.

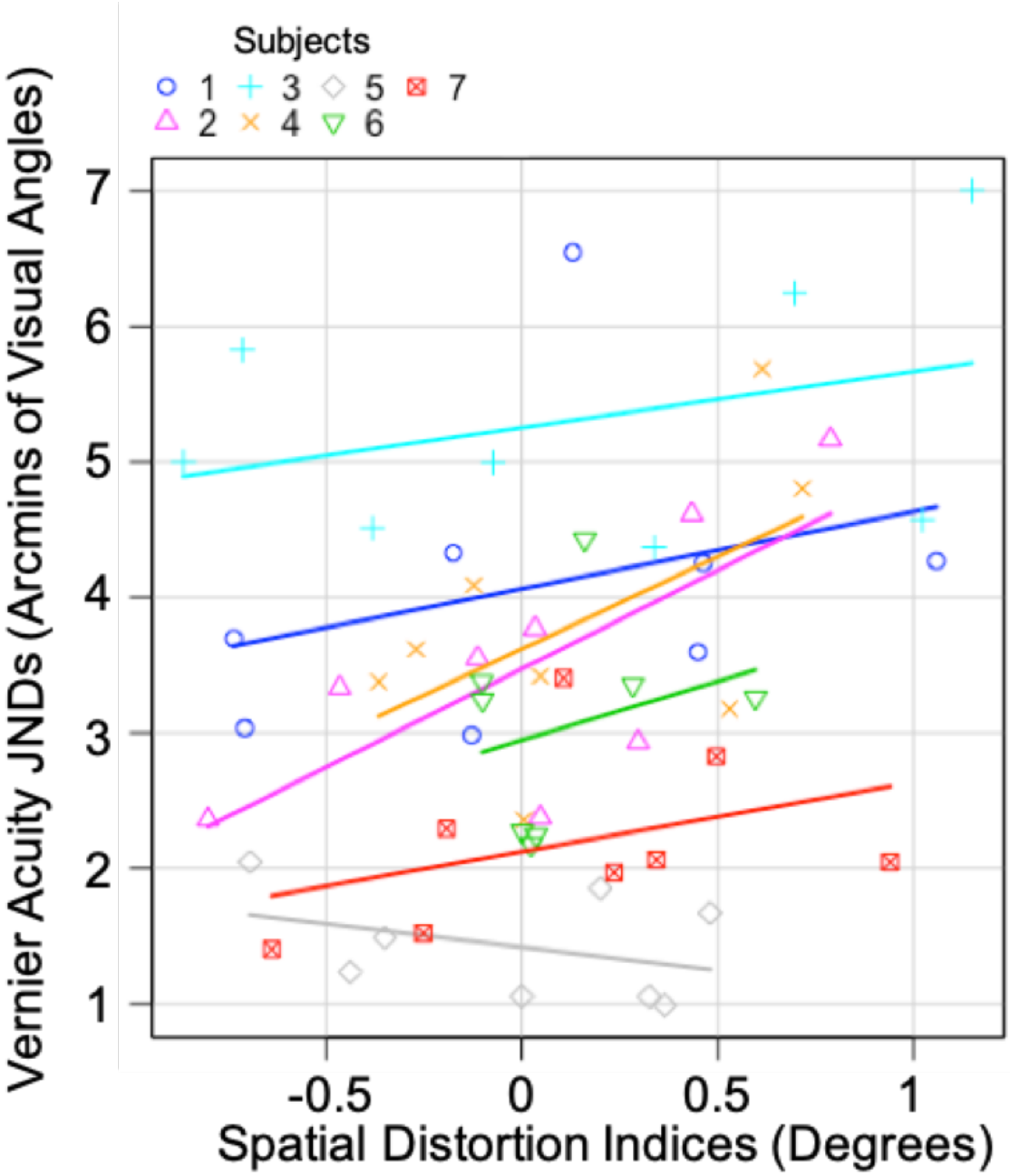

Figure S4: The change of perceived size of the arc stimuli as a function of the angular locations tested for every observer. Subject 1 and Subject 2 are authors. The layout of the angular locations is shown on the bottom right corner.

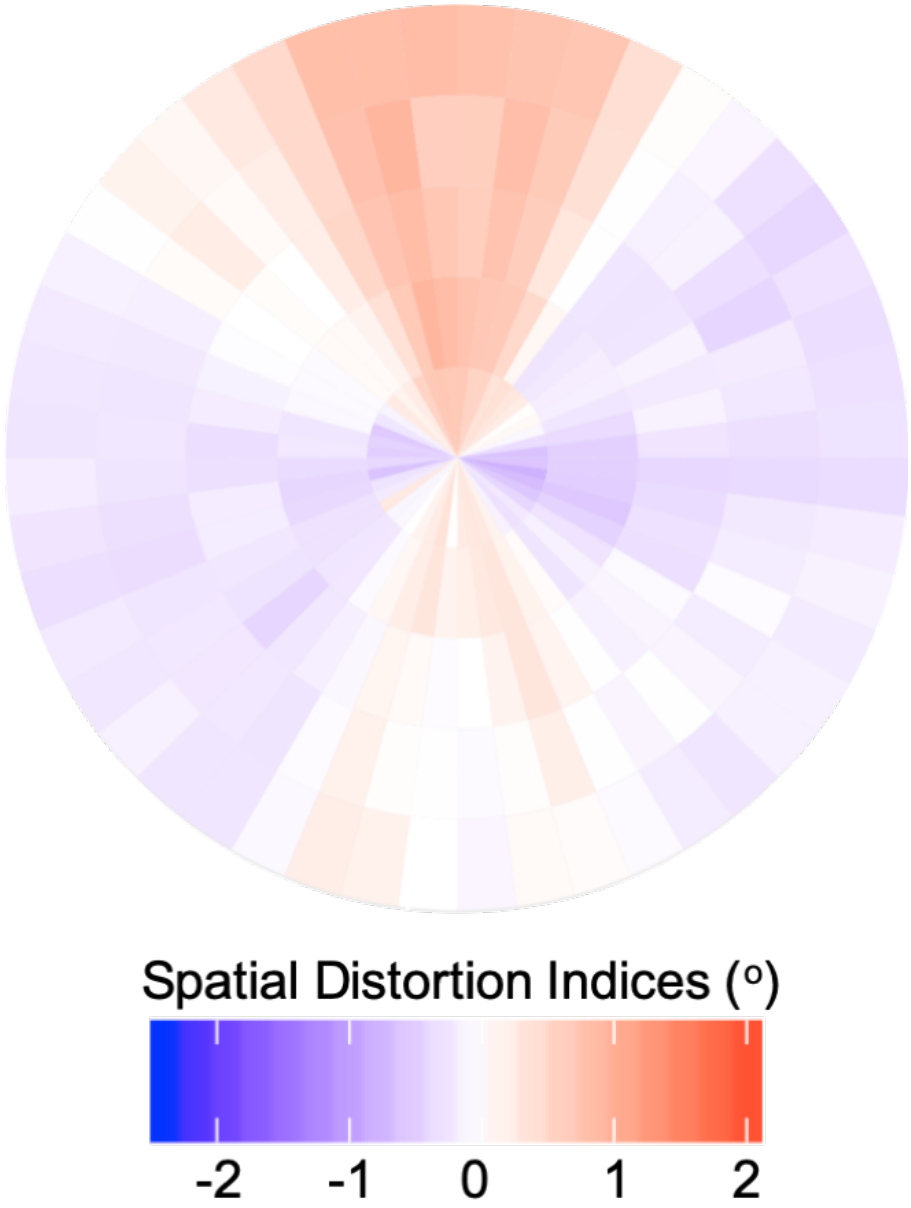

Figure S5: The group mean of the spatial distortion pattern calculated from Chapter 2, Experiment 1 (N=9). Blue area represents visual space compression and red represents expansion. The group average effect is notable and also comports with previous findings on various spatial biases or spatial anisotropies of visual performance (e.g., Low, 1943; Abrams et al., 2012), but it is smaller and easily washed out by the idiosyncratic differences found at the individual-subject level (Fig. 2.1b).

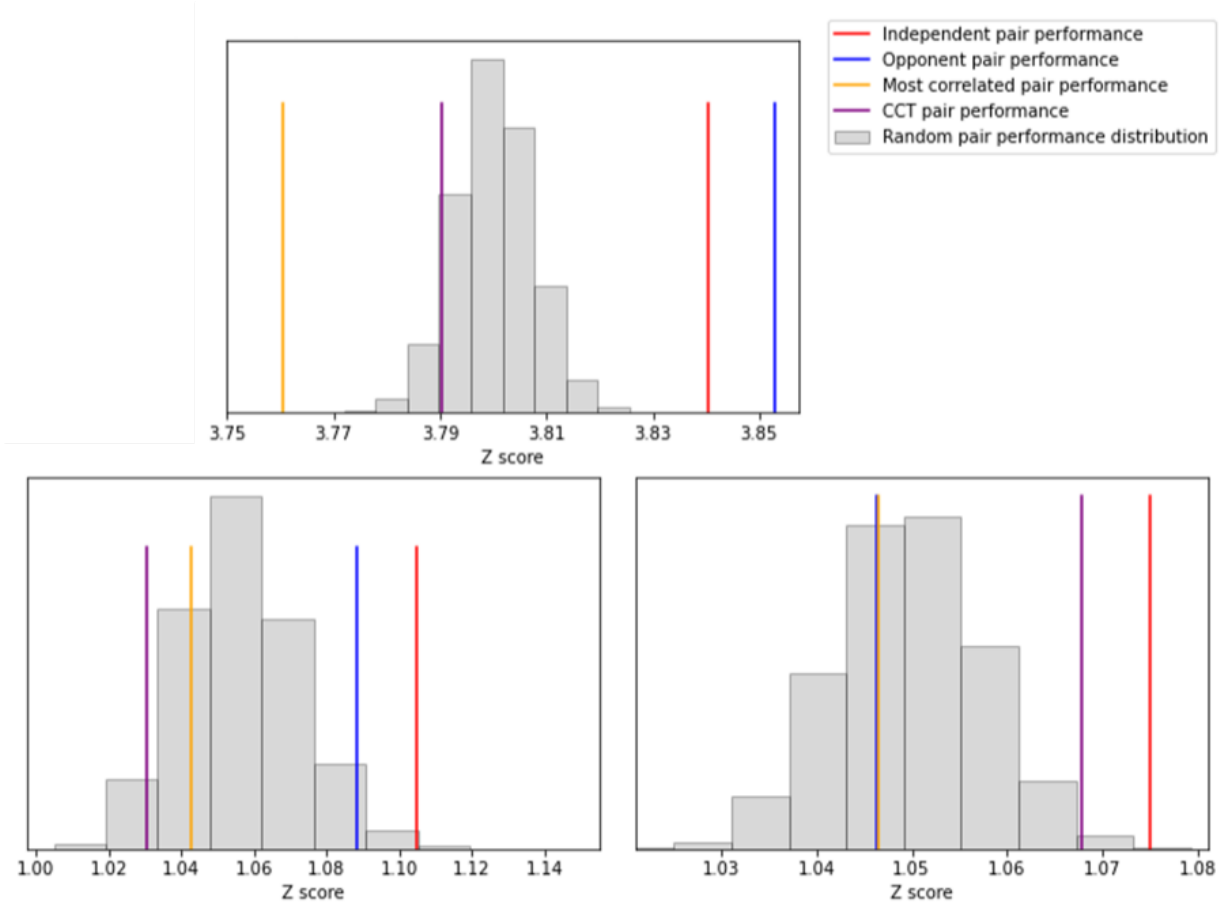

Figure S6: We tested an alternative way of grouping by combining the responses from observers with the most opposite CCT scores calculated based on the cultural consensus theory (CCT; Weller, 2007). Results for localization, shape recognition and emotion tracking tasks are shown in the upper panel, lower left panel and lower right panel respectively. Results showed that unlike independent pairs, CCT pairs were not significantly different from random pairs in any of the three tasks ($p = 0.17$, $p = 0.09$, $p = 0.02$; Bonferroni corrected $\alpha_B = 0.01$). Therefore, we believed that this suggested that our findings is different by other consensus-searching techniques such as CCT.