Flinders
UNIVERSITY
ADELAIDE · SOUTH AUSTRALIA

Collaborative Visual Search

by

Alison Enright

Master of Education (Cognitive Psychology & Educational Practice)
Bachelor Behaviour Science (Honours)
Bachelor of Arts (Psychology)

Thesis
Submitted to Flinders University
for the degree of

Doctor of Philosophy

College of Education, Psychology & Social Work
January 31st 2019

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

This thesis shows work that largely aims to better understand collaborative visual search. More specifically, we aimed to replicate previous findings of collaborative search performance, that is, performance that meets or exceeds the predictions of a uniform weighting model of information integration and extend them to a signal detection task using naturalistic stimuli. We also aimed to investigate a collaborative search strategy when little target information is available to observers.

In the first study, searchers performed a simulated baggage screening task and attempted to detect a target (one of a possible 5 knives) in x-ray baggage images. In Experiment 1, single observers completed the task in separate testing rooms, and teams collaborated in the same testing room. In Experiment 2, single observers and teams completed the task in the same testing room. In Experiment 3, both single observers and teams completed the search task in separate testing rooms. In Experiment 4, finally, single observers and teams were collocated, and stimuli presentation time was fixed (3s). In all four experiments, teams outperformed single observers and achieved sensitivity levels roughly midway between the predictions of the two versions of the uniform weighting model. A meta-analysis using the data from all four experiments confirmed this pattern of results.

In the second study, observers performed a visual search task framed as a medical image reading task and attempted to locate an ‘abnormal cell’ amongst other normal ‘cells’. Top-down target information was limited by using dot-distortion stimuli. In Experiment 1, single observers completed the task in separate testing rooms and team collaborated in the same testing room, whereas in Experiment 2, both single observers and teams completed the search task in separate testing rooms. Teams outperformed single observers in both experiments and collaborative sensitivity again fell in between the predictions of the two versions of the uniform weighting model.

The most consistent finding in both Studies 1 and 2 is that collaborative searchers outperform single seachers. Some of our findings show that teams can even outperform what is expected given their individual sensitivity levels and the similarity between team members’ judgments. Such findings suggest that teams might adopt visual search strategies that work to decorrelate their judgments, resulting in a larger collaborative benefit when integrating their judgments. Another implication of our findings is that non-collocated teams can perform similarly to collocated teams. Finally, we provide evidence that collaboration under conditions of limited target information is valuable.

DECLARATION

I certify that this thesis does not incorporate without acknowledgment any material previously submitted for a degree or diploma in any university; and to the best of my knowledge and belief, does not contain any material previously published or written by another person except where due reference is made in the text.

a. Enright

Friday, 22 June 2018

Alison Enright

LIST OF MANUSCRIPTS

Research from this thesis is currently available at PsyArXiv Preprints:

Enright, A., & McCarley, J. S. (2018, May 16). Collaborative Search in a Mock Baggage Screening Task: A Bayesian Hierarchical Analysis. https://doi.org/10.17605/OSF.IO/8975X

Enright, A., & McCarley, J. S. (2018, June 21). Collaborative searchers outperform individuals in the absence of precise target information. https://doi.org/10.17605/OSF.IO/V6CQK

LIST OF CONFERENCE ABSTRACTS AND PRESENTATIONS

In addition, research from this thesis was presented at:

[1]Enright, A., & McCarley, J. S. (2017). *Collaborative Search: A Bayesian Hierarchical Analysis*. Paper presented at the 2017 Human Factors and Ergonomics Society International Annual Meeting (HFES). Austin, Texas.

Enright, A., & McCarley, J. S. (2017). *Collaborative Search using Naturalistic Stimuli: A Bayesian Hierarchical Analysis*. Paper presented at the 44th annual Australasian Experimental Psychology Conference (EPC). Shoal Bay, Australia.

Enright, A., & McCarley, J. S. (2016). *Differential Carry-Over Effects of Collaborative Visual Search: Benefits Require Practice.* Paper to be presented at the 2016 International Meeting of the Psychonomic Society. Grenada, Spain.

Enright, A., & McCarley, J. S. (2015). *Collaborative Visual Search: Benchmarking Observed Performance Against Models of Optimality.* Paper presented at the 2015 Human Factors and Ergonomics Society International Annual Meeting (HFES). Los Angeles, United States.

Enright, A., & McCarley, J. S. (2015). *Collaborative Searchers Rely on the Most Accurate Observer's Responses in a Joint Decision*. Paper presented at the 42nd annual Australasian Experimental Psychology Conference (EPC). Sydney, Australia.

[1] All of the references noted here originated under Alison Enright's former name, Alison Simpson.

ACKNOWLEDGMENTS

Growing up, my parents enforced one rule above all others – you work, or you study. Clearly, I chose to study.

Studying has been my home, place of comfort and security. I am entirely convinced that this sense of belonging is, at least in part, due to the people who have shared this space with me. My supervisor, Jason McCarley, guided and encouraged me to accomplish an exceptional standard of work. Your time and advice are absolutely dripping with wisdom and this thesis simply would not have been possible without you. Thank you.

In the second year of my candidature, the Brain and Cognition Lab adopted me. Thank you, Nicole, for initiating this process, showing me the ropes, and helping me get my first paper published. My B & C comrades produced some of my fondest memories over the past four years. Thank you, Nathan, Dan, Ellie, Steph, Bek, Megan, Owen, Oren, and Scott for contributing to hours of venting, laughing, and lunching. Mike, your guidance is second only to your spag bog crown. Thank you for welcoming me into your lab and supporting me to attend conferences. Liz, I will always remember our chats, tears, and celebrations.

Multiple significant life events occurred while completing this doctorate. Despite all that life planned, my family showed me unwavering support. Thanks to my parents for making this an option, it certainly would not have been without your help. To the best partner an academic could hope for, Froggy, our girls would be lost somewhere in the woods, surviving off the land if it weren't for you. Thank you for taking care of all of us. And to my girls, you will only know the strength you have inside you when you have to ask your family to wait while you work on a dream. Thank you for waiting. This is for you.

This achievement is not only mine but reflects the combined efforts of those who supported me. To all of you, thank you. I guess it's time to go to work now.

This work was supported by an Australian Government Research Training Program Scholarship.

CHAPTER 1:

LITERATURE REVIEW

A key component of everyday visual behavior is the ability to find important items in a scene filled with distracting items. Visual search is a nontrivial activity and has been the focus of intense interest to visual cognition (Duncan & Humphreys, 1989; Itti & Koch, 2000; Treisman & Gelade, 1980; Wolfe, 1994) and human factors (e.g., Beck, Lohrenz, & Trafton, 2010; Wickens, Alexander, Ambinder, & Martens, 2004). Traditionally, search has been studied by examining the performance of single observers. However, visual search is often collaborative, in daily life, such as when a couple attempt to locate a street address while driving, and in high-stakes domains such as transportation security officers jointly inspecting the same x-ray baggage image.

Signal detection theory

Collaboration during visual search requires multiple observers to reach a joint decision. Signal detection theory (SDT; Green & Swets, 1966; Macmillan & Creelman, 2005) provides models of how groups can reach decisions from probabilistic evidence distributions, and a framework for predicting and assessing collaborative search performance (Bahrami et al., 2010; 2012; Sorkin & Dai, 1994; Sorkin, Hays, & West, 2001). Conventional signal detection tasks require observers to judge on each trial if stimuli were drawn from a signal-plus-noise distribution or a noise-alone distribution (Green & Swets, 1966; MacMillan & Creelman, 2005, Tanner & Swets, 1954). In a study of transportation security screening, for example, a task might present signal-plus-noise trials that contain a threatening item (signal stimulus) embedded in a bag cluttered with non-threatening items (noise stimuli), and noise-alone trials that present only the bag of non-threatening items. On each trial, the observer reports whether or not a signal was detected.

Table 1 shows the four possible response outcomes for a trial in a yes/no SDT task, i.e., an SDT task that requires a ‘yes/target present’ or ‘no/target absent’ response. On a signal-plus-noise trial, a correct response is referred to as a *hit* and an incorrect response is referred to as a *miss*. An accurate response on a noise-alone trial constitutes a *correct rejection*, and an inaccurate response on a noise-alone trial constitutes a *false alarm*.

Table 1.1. Four response outcomes for a trial of a yes/no SDT task

	Response	
Stimulus	"yes/target present"	"no/target absent"
Signal (*S*)	hit	miss
Noise (*N*)	false alarm	correct rejection

Observers' responses are based on the sampled value of an internal *decision variable* (Stanislaw & Todorov, 1999) that corresponds to the evidence for or against the signal each trial (Green & Swets, 1966, Metz & Shen, 1992). As its name implies, the decision variable is non-deterministic; because of fluctuations of observer state and random variability in stimulus properties, strength of evidence for or against a signal varies from trial-to-trial even when holding the actual presence or absence of a signal constant (Burgess & Colborne, 1988). An observer reaches a discrete yes-or-no judgment by comparing the value of the decision variable sampled that trial to a *criterion* value, a threshold determining whether an observer responds positively or negatively (Green & Swets, 1966, Macmillan and Creelman, 2005). If the decision variable exceeds the criterion value, the observer responds 'yes'. Otherwise, she responds 'no'.

Figure 1 shows the distribution of the potential values realized by the decision variable across signal-plus-noise and noise-alone trials, as well as the regions of the distributions corresponding to the four possible response outcomes for a yes-no task. The figure assumes a standard equal-variance Gaussian model, that is, normally distributed signal-plus-noise and noise-alone distributions with the same standard deviation but different means (Macmillan & Creelman, 2005). The signal distribution (right curve) represents signal-plus-noise trials and the noise distribution (left curve) represents noise-alone trials. The means of the noise and signal distributions are 0 and μ_S respectively.

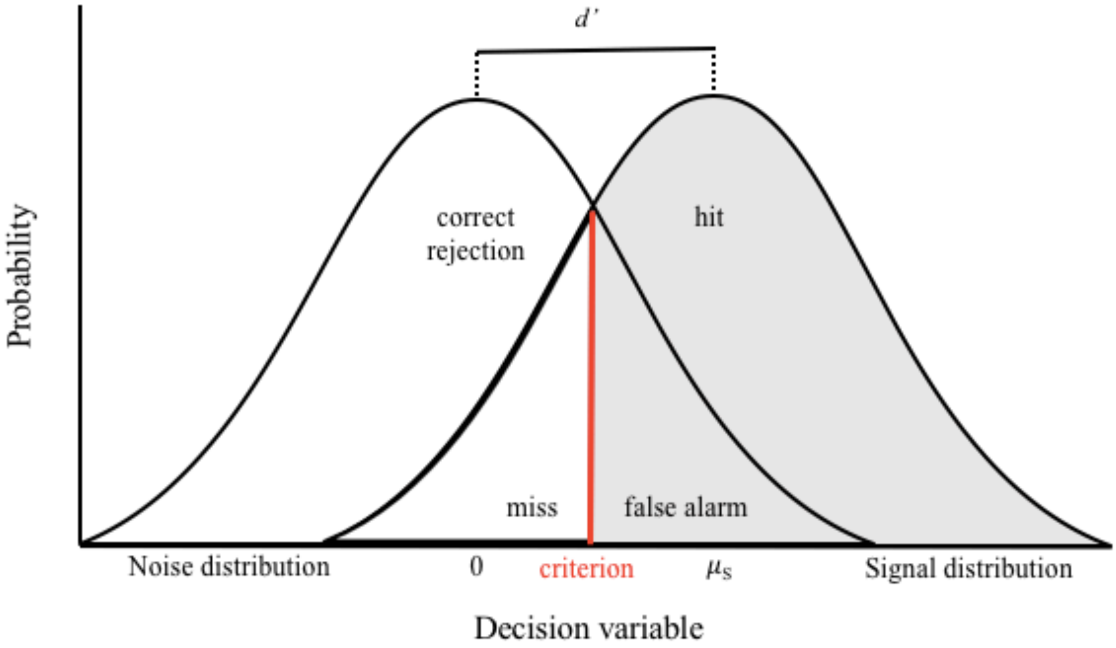

Figure 1.1. Distribution of decision variables for signal and noise trails in a yes/no task showing the four response outcomes, criterion, and *d'*.

Overlap between the signal and noise distributions determines an observer's *sensitivity*, ability to discriminate between signal-plus-noise and noise-alone stimuli (Green & Swets, 1966; Macmillan & Creelman, 2005; Tanner & Swets, 1954). A perfectly sensitive observer presents a hit rate of 1 and false-alarm rate of 0. A perfectly insensitive observer, though, is unable to discriminate between the two stimuli types (i.e., signal and noise) and presents the same probability for both hit and false-alarm rates. In figure 1.1, the hit rate equals the proportion of the signal distribution that exceeds the criterion value (e.g., the red line), whereas the false-alarm rate is the proportion of the noise distribution that exceeds the criterion value.

The hit and false-alarm rates shift according to observers' *response bias*, tendency to respond positively or negatively (Green & Swets, 1966; Macmillan & Creelman, 2005). Response bias is determined by the criterion placement. An unbiased criterion value produces equal hit and correct rejection rates. A *liberal* criterion is lower, increasing the hit and also the false-alarm rates, and a *conservative* criterion is more stringent, decreasing the hit and false-alarm rates.

Performance measures

Sensitivity. Conventionally, sensitivity is measured using *d'* (Green & Swets, 1966; Macmillan & Creelman, 2005),

$$d' = z(H) - z(F),$$

where z is the inverse normal transformation and H and F are the hit and false alarm rates, respectively. More specifically, *d'* is the distance between the means of the signal and noise distributions, in standard deviation units. *d'* is a useful measure of sensitivity because it is robust against changes in response bias. However, this is true only if the signal and noise evidence distributions are normal and share the same standard deviation. If either assumption is violated, *d'* will vary with response bias, regardless if the overlap between the signal and noise evidence distributions remains constant (Stanislaw & Todorov, 1999).

Receiver operating characteristic. A receiver operating characteristic (ROC) or isosensitivity curve plots hit-rate as a function of the false-alarm rate while holding sensitivity constant (Green & Swets, 1966; Macmillan & Creelman, 2005). Figure 1.2 shows a typical ROC. Empirically, the multiple hit and false-alarm rate pairs needed to plot the curve are collected by varying the observer's response criterion while holding task difficulty fixed. This can be done by manipulating signal rate or response payoffs across blocks of trials, or by collecting confidence-rated judgments in a single block of trials (Macmillan & Creelman, 2005).

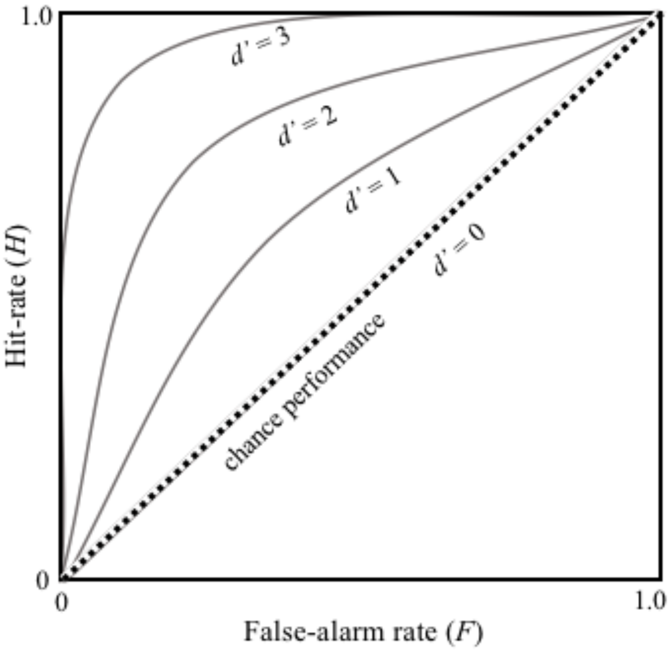

Figure 1.2. Three ROCs of differing sensitivity levels and the chance line (dotted black line) (adapted from Macmillan & Creelman, 2005).

Analysis of receiver operating characteristic curves (ROC; Green & Swets, 1966; Stanislaw & Todorov, 1999) provides alternative measures of sensitivity that do not require the assumption of equal variance in the signal and noise evidence distributions. The major diagonal denotes chance performance, the level at which hit and false-alarm rates are equal. Above-chance sensitivity levels shift the ROC toward the upper left corner, where $H = 1$ and $F = 0$, indicating perfect performance (Macmillan & Creelman, 2005). The slope of the ROC is the change in hit-rate relative to the change in false-alarm rate, and as such, the slope of a ROC curve decreases as responses are biased towards ‘yes’ (Green & Swets, 1966; Macmillan & Creelman, 2005; Tanner & Swets, 1954).

Plotting a ROC in standardized normal space produces a *z*ROC, which can be used to test the assumptions of equal-variance Gaussian evidence distributions. Assume, following convention, that the evidence distribution for noise trials is Gaussian with mean 0 and standard deviation 1, and the signal-plus-noise distribution is Gaussian with mean $\mu_{S+\mathrm{N}}$ and standard deviation $\sigma_{\mathrm{S+N}}$. Then, the *z*ROC is a linear function with *y*-intercept $\mu_{S+\mathrm{N}}$ and slope $\sigma_{\mathrm{S+N}}$,

$$z(H) = \mu_{S+\mathrm{N}}/\sigma_{\mathrm{S+N}} + 1/\sigma_{\mathrm{S+N}} \times z(F).$$

The slope of the zROC equals the inverse of the standard deviation of the signal-plus-noise distribution (Macmillan & Creelman, 2005; Wickens, 2002), and a *y*-intercept equal to the mean of the signal-plus-noise distribution divided by the standard deviation of the signal-plus-noise distribution (Stanislaw & Todorov, 1999). As such, slopes that equal 1 indicate equal variance in the evidence distributions whereas slopes that do not equal 1 indicate unequal variances.

Figure 1.3 shows a family of typical *z*ROCs with slope = 1. When the zROC has unit slope, *d'* is equal to the distance in *z* units from the *z*ROC line to the chance line.

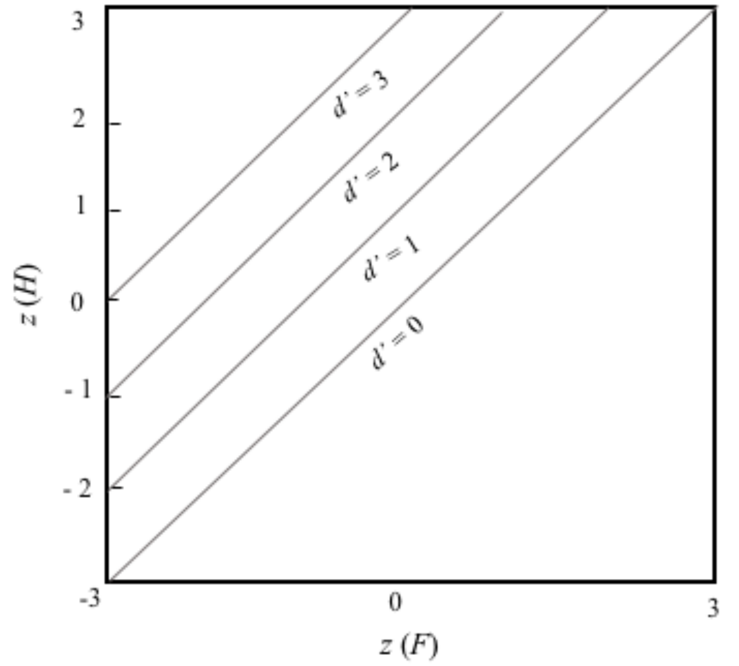

Figure 1.3. ROCs for a standard SDT task on *z* coordinates (adapted from Macmillan & Creelman, 2005).

In the event that the slope of the zROC does not equal 1, the assumption of equal variance evidence distributions is violated, and an alternative to *d'* is required to measure sensitivity. Figure 1.4

presents a zROC with a slope unequal to 1, and three alternative sensitivity measures, d'_1, d'_2, and d'_e. The first, d'_1, is the distance between the ROC and the major diagonal at the point where *z*(*H*) equals 0 and provides a slightly amplified sensitivity measure. The second, d'_2, is the difference between the ROC and the major diagonal but at the point where *z*(*F*) equals 0 and provides a slightly compressed sensitivity measure. The third, d'_e, is based on the mean of the standard deviations of the evidence distributions (Egan, Schulman, & Greenberg, 1959; Macmillan & Creelman, 2005) and is defined as,

$$d'_e = \frac{2\mu_S}{1+\sigma_S}$$

where $\mu_S$and $\sigma_S$are the mean and standard deviation of the signal distribution respectively. When noise-alone and signal-plus-noise distributions are equal-variance, d'_e reduces to d'. Otherwise, d'_e may be above or below d'(Macmillan & Creelman, 2005).

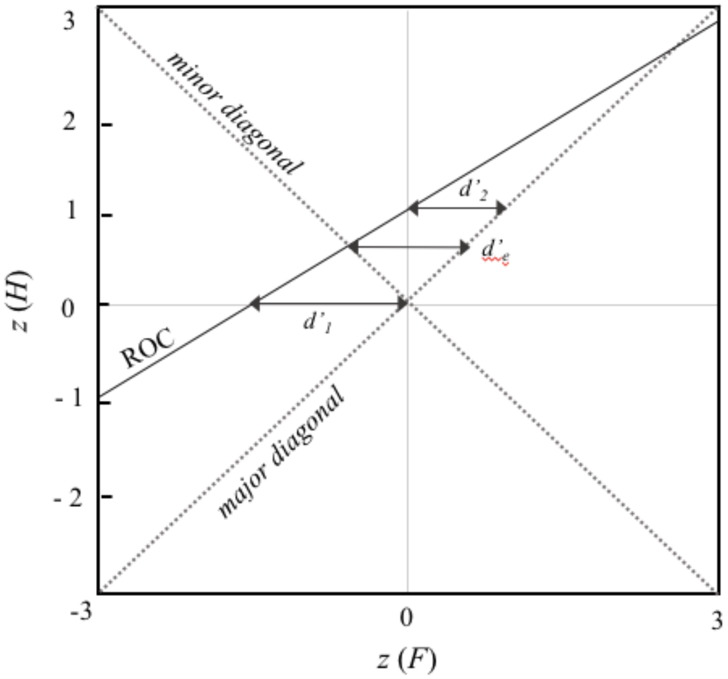

Figure 1.4. zROC showing three measures of sensitivity: d'_1 and d'_2 are dependent on the standard deviation of the signal and noise distributions respectively, and an average of d'_1 and d'_2, d'_e (adapted from Macmillan & Creelman, 2005).

Visual search as a form of Signal Detection

Theories of visual search performance (e.g., Duncan & Humphreys, 1989; Itti & Koch, 2000; Nodine & Kundel, 1987; Treisman & Gelade, 1980; Wolfe, 1994) generally share a basic two-stage structure. In the first stage, the viewer orients to the search field using parallel preattentive channels that register low-level features maps such as colour, shape, movement, and orientation, which generate a saliency map of the visual scene. If the target object is not detected readily, orientation gives way to a stage of slower, serial attentional scanning guided by preattentive outputs.

Bottom-up and top-down activation contribute to differential activation of feature maps (Wolfe, 1994). Bottom-up activation is stimulus-driven and guides attention to the most salient search item. The observer requires no knowledge of the search task, rather salient items appear distinct from neighbouring items, thereby producing greater activation. When featural properties of a search item are not unusual, however, top-down activation is needed. Top-down activation is knowledge-driven and requires the observer to adopt a target template to prioritize to locations likely to contain target featural properties (Wolfe, 1994).

In a visual search task requiring only parallel processing, the target produces the greatest level of activation in the activation map, regardless of how many additional stimulus items are presented with the target, and often appears to ‘pop-out’ of the stimulus display, capturing attention (Treisman, 1985; Treisman & Gelade, 1980; Wolfe, 1994). In a visual search task requiring serial processing, attention is directed to the most highly activated location on the activation map and continues to move from one location of high activation to another, in order of decreasing activation, until the target is located, or the search is terminated (Duncan & Humphreys, 1989; Itti & Koch, 2000; Nodine & Kundel, 1987; Treisman & Gelade, 1980; Wolfe, 1994).

Psychophysics of visual search. Signal detection theory is not the only feasible model of visual search. Palmer et al. (2000) note that the above two-stage theories of visual search performance (e.g., Duncan & Humphreys, 1989; Itti & Koch, 2000; Nodine & Kundel, 1987; Treisman & Gelade, 1980; Wolfe, 1994) assume a high threshold model of perceptual judgments.

The classical feature of high threshold theories is that a ‘target present’ response, when no signal is in fact present, is purely the result of an observer guessing. In other words, noise stimuli cannot produce a target-present internal state (Macmillan & Creelman, 2005; Palmer et al., 2000). This is in contrast to low threshold theories such as SDT, which allow the possibility that a noise stimulus produces a target-present internal state. (Note that various authors define low-threshold theory differently. Whereas Palmer et al. (2000) use the term to denote any model in which a noise stimulus can be confused with a signal, others use it more specifically to denote models assuming discrete pre-decisional mental states; see Macmillan and Creelman, 2005. The current discussion adheres to Palmer et al.’s usage.)

Palmer et al., (2000) compared six visual empirical search phenomena – the effects of set size, multiple targets, distractor heterogeneity, target-distractor discriminability, response bias, and external noise – to the predictions of high and low threshold theories. They concluded that the empirical results for all of the six were consistent with the predictions of low threshold theories and inconsistent with those from high threshold theories. Similarly, Eckstein (1998) and Eckstein, Thomas, Palmer, and Shimozaki (2000) found that signal detection theory models accurately predicted differences in the efficiency of feature and conjunction search, without appealing to a two-stage, parallel-to-serial processing model. Additionally, Verghese’s (2001) review of visual search data provides considerable evidence that at least some visual search results can be explained without a second-stage, limited-capacity feature, rather a signal detection explanation in which the second stage is a decision rule. Discussion continues as to whether visual data are better reconciled with a two-stage model or a single-stage parallel model built on signal detection theory (Palmer, Fencsik, Flusberg, Horowitz, & Wolfe, 2011; Verghese, 2001). Nonetheless, findings from Palmer et al. (2000) and others (Verghese, 2001) demonstrate at the very least that visual search performance is amenable to analysis with the methods of signal detection theory.

Collaborative visual search

Intuition, as captured in the aphorism, “two heads are better than one,” suggests that teams of collaborators ought to outperform individuals in search tasks, as in fact they do in many other cognitive tasks (Kerr & Tindale, 2004; Laughlin, Bonner, & Miner, 2002; Levine & Moreland, 1990). The benefits of collaboration, though, are often modest, group members integrating biased information when reaching a joint decision, for example, can reach decisions based on subsets of information that fail to take into account other, non-biased information, thereby producing degraded collaborative decisions that are not reflective of all potentially pooled information (Stasser & Titus, 1985). Such effects give little assurance that teams will generally outperform individuals in visual search tasks (Kerr & Tindale, 2004).

The belief that groups generally outperform individuals likely comes from strong mathematical arguments showing the benefit of aggregated responses. Marquis de Condorcet’s jury theorem (Condorcet, 1785) was among the first to mathematically demonstrate that a ‘majority rule’ of aggregated votes from individuals produces near-perfect accuracy as long as the number of individuals is sufficiently large. Galton (1907) provided empirical support for the value of information aggregation when he asked local livestock fair goers to guess the weight of a specific ox that had been slaughtered and dressed. Galton (1907) found that the aggregated guesses of 800 fair goers, with no butchery expertise, produced a more accurate judgment than the best expert.

Research in this ‘wisdom of crowds’ effect (Suriowiecki, 2004) resulted in a number of models of collaborative decision making (Bahrami et al., 2010, 2012; Davis, 1996; Hastie & Kameda, 2005; Hill, 1982; Kameda, Tsukasaki, Hastie, & Berg, 2011; Sorkin & Dai, 1994; Sorkin, Hays, & West, 2001; Sorkin, West, & Robinson, 1998), positing various methods by which individuals might aggregate their judgments.

Condorcet group models. Sorkin et al., (1998) referred to groups that rely on some form of a majority rule of aggregated individual votes as Condorcet groups. Following Condorcet's (1785) theorem, Condorcet group models base collaborative decisions on the likelihood that individual group members produce correct responses is above .5. If so, then the probability of a correct response from a majority response will increase with group size (Sorkin et al., 1998).

Figure 1.5 presents the signal detection system of a Condorcet group composed of *m* members, each of which with their own sensitivity level, *d'* (Sorkin & Dai, 1994; Sorkin et al., 2001; Sorkin et al., 1998). Each group member receives an input of either a signal or noise. After observing the input stimuli, group members estimate the signal's likelihood, *X*. If a group member's estimate exceeds their response criterion, *c*, then that member responds 'yes', signal present. If a group member's estimate falls below their response criterion, the member responds 'no', signal absent.

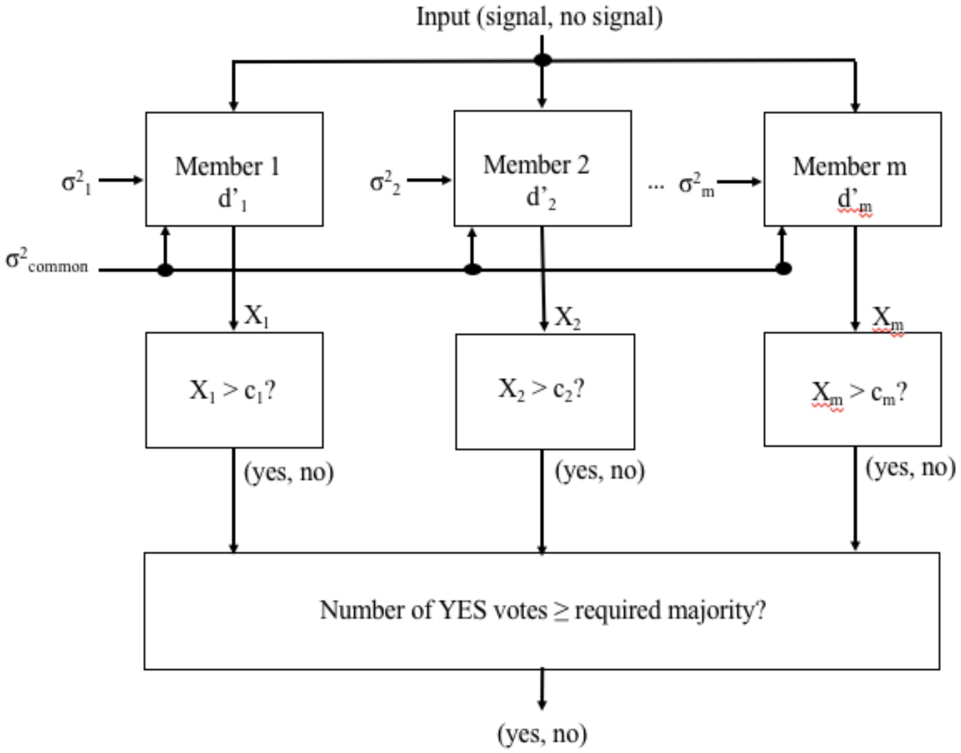

Figure 1.5. Signal detection of a Condorcet group where d' is the sensitivity index, c indicates response criterion, and X and is the group member's estimate of the signal's likelihood (adapted from Sorkin et al., 1998).

Two noise sources contribute to the decision process. Unique noise, σ^2_i, $i = 1, 2 \ldots m$ is the variance isolated to an individual team member, and common noise, σ^2_{common}, is variance shared among team members. For example, consider a pair of transportation security officers jointly inspecting a baggage x-ray. Unique noise might arise from differences in the searchers' internal sensory noise, or from differences in their search patterns. Common noise might arise from ambiguities or degradations in the x-ray image itself, or from similarities in the viewers' search patterns.

The collaborative decision is determined by the majority rule applied to individual group members' decisions (Sorkin et al., 1998). Majority rules range from liberal, the least number of group members required to reach a consensus, to strict, the largest number of group members required to reach a consensus. The *simple majority* rule requires 'yes' responses from at least half of the group members (*m*/2). The *unanimous majority* rule requires 'yes' responses from all group members (*m*). A number of intermediate majority rules exist within the extremes of the simple and unanimous majority rules, for example, a rule that requires 'yes' responses from all group members except for one (*m* – 1 rule; Sorkin et al., 2001). Analyzing ROCs, Sorkin and colleagues (1998) showed that collaborative performance improved with group size, especially for groups employing a more lenient majority rule, e.g., *m*/2. Collaborative performance was poorest for groups using stricter majority rules, i.e., *m*, *m* – 1 (Sorkin et al., 1998).

But regardless of the majority rule applied, Condorcet groups perform well below the predictions of other models of collaborative decision making (Bahrami et al., 2010; Sorkin & Dai, 1994; Sorkin et al., 2001). One explanation for the inefficiency of Condorcet groups is that they shrink continuous judgements into binary responses and then base collaborative decisions on the unweighted combination of these binary responses (Sorkin et al., 2001). In addition to essentially eliminating useful information in group members' continuous responses, weighing each member's response similarly results in further information loss because the responses from more competent members are essentially discounted and treated the same as the responses from the least competent members. As such, Sorkin et al. (2001) treats the level of sensitivity achieved by Condorcet groups as an approximate lower bound of collaborative efficiency.

Models of group signal detection. Where the collaborative performance levels achieved by Condorcet groups represent the lower bound of predicted collaborative performance, models of group signal detection incorporating more sophisticated information integration strategies establish the upper bounds of collaborative performance (Sorkin & Dai, 1994; Sorkin et al., 2001).The group decision process again begins with a signal-plus-noise or noise-alone event from which group members sample information to reach individual judgments (see figure 1.6). To reach a joint decision, group members again combine their judgements in some fashion.

Ideal collaborative performance comes from the *Optimal Weighting model* (OW; Bahrami et al., 2010; Sorkin & Dai, 1994; Sorkin et al., 2001). The OW model assumes that individual team members' judgments are averaged in a weighted manner to produce a team decision variable. The value of a team's optimal decision variable, X_{team}, is given by,

$$X_{team} = \sum X_i d'_i$$

where X_i is the decision variable for team member i, and d'_i is team member i's sensitivity.

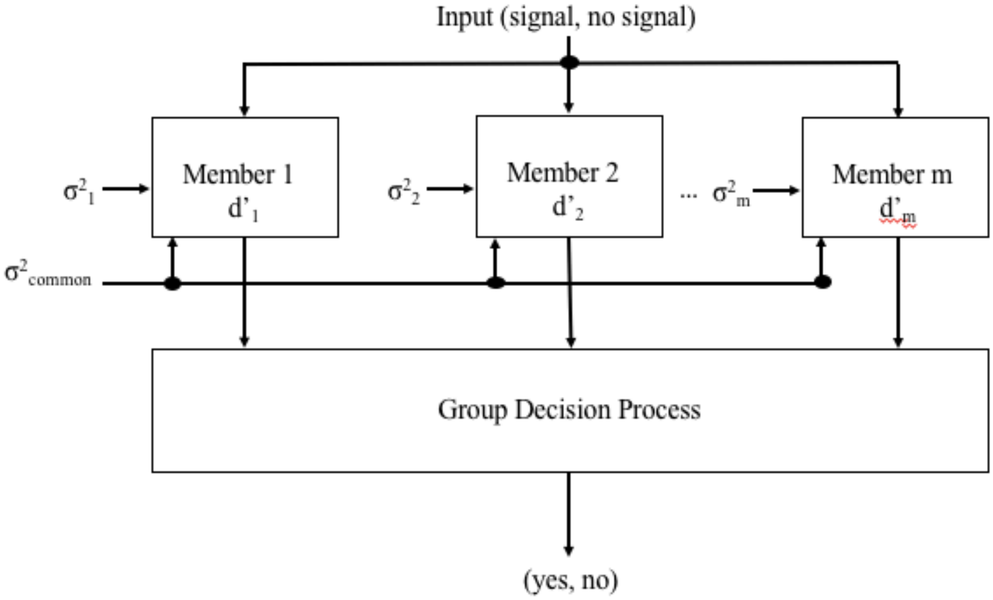

Figure 1.6. The general group signal detection paradigm. This figure was adapted from Sorkin et al. (1998) and Sorkin et al. (2001).

Assuming that the team members' individual judgments are stochastically independent, team sensitivity under the OW model is,

$$d'_{OW} = \sqrt{\sum d'^2_i}.$$

The level of sensitivity achieved by the ideal group provides the upper bound on collaborative performance (Sorkin et al., 2001). When team members present correlated judgments, collaborative sensitivity decreases,

$$d'_{\text{correlated OW}} = \sqrt{m}\ \left[\left(\frac{\mathrm{Var}(d')}{1-\rho} + \frac{[\mathrm{mean}(d')]^2}{1+\rho(m-1)}\right)\right]^{1/2},$$

where m is the total number of group members, $\mathrm{Var}(d')$ is the variance of the individual team members' d' values, and ρ is the correlation among members' judgments. However, to use this formula, the correlation needs to be known, which is often challenging, given that the group members individual decision variables are internal and not directly observable (Metz & Shen, 1992).

The value of collaboration decreases when team members produce correlated judgments because team members share redundant information, limiting their capacity to make use of novel information in the collaborative decision-making process. Team members' responses become correlated when common noise sources, such as those noted in the transportation security officers example above, increase. It seems unlikely that a pair of observers could overcome such common noise sources, however Sorkin et al. (2001) note that *ideal observers* ought to.

The Uniform Weighting Model (UW; Sorkin & Dai, 1994) is similar to the OW model, but assumes that individual team members' signal likelihood judgements are weighted equally when they are averaged to reach a group judgement. Assuming an equal-variance Gaussian model again, group d' under the UW model is,

$$d'_{UW} = \frac{\sum d'_i}{\sqrt{N}}\,.$$

Bahrami et al. (2010) note that uniform weighting is similar to a circumstance in which team members do not directly communicate signal likelihood estimates, but instead convert an internal evidence representation into a confidence rating through a standard normal transformation. During trials, this is straight forward – participants respond according to their unique confidence estimates about target presence. Bahrami and colleagues argue that a more complex operation occurs to develop that sense of confidence, though, by transforming information from the normally distributed evidence to a z-score that operates as a yardstick for their internal confidence estimates. When collaborators are equally sensitive, the UW model is equivalent to the OW model. Otherwise, the unweighted model produces lower sensitivity.

The *Best Decides Model* (BD; Bahrami et al., 2010; Denkiewicz, Rączaszek-Leonardi, Migdal, & Plewczynski, 2013) predicts less efficient collaborative performance than either the OW and UW models, and holds that a team reaches its decision by deferring to the most sensitive member's judgment each trial. Sensitivity for the team thus equals that of the most sensitive team member,

$$d'_{\mathrm{BD}} = \max(d'_1, d'_2).$$

Empirical evidence indicates collaborative performance in some perceptual and statistical decision-making tasks approaches levels predicted by the UW model (Bahrami et al., 2010; Sorkin et al., 2001). Bahrami et al. (2010) presented dyads with two viewing intervals each containing six Gabor patches. A single target, a patch with slightly elevated contrast, was present in either the first or second viewing interval, and the participants' task was to decide which interval contained it. Dyad members first provided individual responses, without consulting each other, and in the event their responses conflicted, a joint response was requested. Trial feedback was then communicated onscreen to dyad members. In Experiment 1, stimuli were approximately equally discriminable for all participants. In Experiment 2, visual noise was added to the stimuli for one participant within each dyad in order to elicit different levels of sensitivity for the two dyad members. In Experiment 3, communication was restricted to their yes or no judgements, and in Experiment 4, no feedback was provided. Results from both Experiments 1 and 2 were consistent with the UW model. Dyads in the third experiment produced collaborative performance no better than the more sensitive observer, indicating that feedback alone was insufficient to give collaborative benefit. Collaborative sensitivity in Experiment 4 was similar to that predicted by the UW model indicating that feedback was unnecessary to produce collaborative benefit, in turn suggesting that communication plays a key role in collaborative benefit.

Using a similar paradigm, Bahrami et al. (2012) likewise found evidence for collaborative performance levels predicted by the UW model. The participants' task in this study was identical to that of Bahrami et al. (2010)'s experiment. In Experiment 1, different sensitivity levels for dyad members were elicited by including visual noise to the stimuli for one team member and participants communicated verbally or non-verbally (i.e., via confidence visual schema). In Experiment 2, participants viewed identical stimuli and communicated verbally, verbally and non-verbally, or not at all. The data from Experiment 1 indicated that when teams communicated verbally, collaborative sensitivity was worse than that of the better observer, consistent with UW model predictions. However, UW model predictions slightly overestimated collaborative performance of verbally communicating teams. Non-verbally communicating teams, though, performed no worse than the better observer, and

UW model predictions slightly underestimated collaborative performance. Bahrami et al. (2012) attributed the collective failure in verbally communicating groups to individuals' underlying cognitive biases. Data from Experiment 2 showed that verbally communicating teams outperformed the better observer and matched UW model predictions. When teams communicated with strictly non-verbal or no communication, however, collaborative sensitivity was not better than that of the better observer, and UW model predictions overestimated empirical collaborative sensitivity. Taken together, the findings from Experiment 1 and 2 suggest that the type of communication should be dependent on the similarity of observers' sensitivity levels, such that when observers present with dissimilar sensitivity levels, direct verbal communication degrades the value of collaboration. However, when observers present with similar sensitivity levels, direct communication should be used.

Using Bahrami et al.'s (2012) data (from Experiment 2 – verbally and non-verbally communicating teams described above) Bang et al. (2014) explored whether teams with dissimilar sensitivity levels could achieve UW-level performance by using team members' confidence levels or the judgment of the team member with the fastest reaction time (i.e., justified by the inverse relationship between confidence and response time). Data were submitted to two algorithms, the maximum confidence slating (MCS) and the minimum reaction time slating (MRTS), whose output responses were compared to teams' observed responses. Results indicated that sharing confidence estimates produced UW-level performance only when team members shared similar sensitivity levels and could interact (verbally communicating teams).

Sorkin, Hays, and West (2001) also assessed collaborative decision-making performance, manipulating task difficulty. Participants performed a signal detection task individually and in groups ranging in size from 2–7 members. On each trial, participants were presented with a display of nine analog gauges. Values of the gauges were drawn from one of two equal variance Gaussian distributions, one of higher mean than the other, and the participants' task was to decide which distribution the values were drawn from. Task difficulty was manipulated by changing the display signal-to-noise ratio (DSNR), the standard deviation of the evidence distributions. In Experiment 1, the distribution of DSNR was equal for all group members, equal task difficulty, or different for half of the group members, unequal task difficulty. Stimulus displays were also manipulated such that they were either independent across group members ($\rho = 0$) or partially correlated ($\rho = 0.25$). Groups with independent displays performed better than groups with correlated displays and better than individuals, but with efficiency that was less than predicted by the either OW model or a simple majority rule model, i.e., *m/2*. Deviations from optimal group performance did not seem to result from inefficient combination of the individual participants' decisions. Rather, group performance was poorer than expected because individual observers put forth less effort into the task as group size increased (shown by decomposing a measure of group efficiency into two components – group member effort and group decision aggregation efficiency – and comparing both components' contribution to the overall group detection performance).

In Sorkin et al.'s (2001) second experiment, individual sensitivities were measured by requiring participants to make a response prior to the collaborative decision process. This allowed the researchers to more conclusively rule out inappropriate weighting strategies as a cause of the inefficiencies in collaboration in Experiment 1. Furthermore, by collecting individual responses prior to the group response, the researchers were able to calculate the correlation between group members' judgments. These initial individual responses were presented onscreen and available to group members during the group deliberation. All group members in Experiment 2 viewed displays of the same DSNR. The results of Experiment 2 showed no evidence of correlations between group members' individual responses, a result that was as expected given that displays were generated independently for each team member. As in Experiment 1, groups performed better than individuals, but collaborative performance decreased with group size. Analyses confirmed that the decreasing collaborative performance was not a result of inappropriate weighting strategies, but of decreasing individual group member effort.

Other evidence, from Malcolmson, Reynolds, and Smilek (2007), compared the collaborative performance of empirical and nominal teams in a visual search task. In their study, two participants completed the task in the same testing room sharing one computer (empirical teams) and independently in separate testing rooms (nominal teams). Trial order was identical for the two participants in the nominal teams so that individual judgments could be combined to reach one team judgment. Specifically, target present responses required one or both team members responded target present and target absent responses required both team members responded target absent. Empirical teams were encouraged to devise a search strategy that took advantage of the fact that there were two observers performing the task. This encouragement was reiterated every 40 trials, and teams were asked to check how their strategy was working. Participants comprising nominal teams were encouraged to evaluate their independent strategies. Results indicated that empirical teams outperformed nominal teams (Experiment 1). Empirical teams reported that the search strategy they employed was to divide the display into halves, making each team member responsible for one half. Malcolmson et al. (2007) tested whether collaborative performance above that of nominal groups could be explained by social facilitation effects. Experiment 2 replicated the same pattern of results, indicating that social facilitation effects were unlikely to contribute to the obtained collaborative gain.

Social processes in collaborative performance. Groups sharing similar or redundant information is the heart of the 'hidden profile' procedure (Stasser & Titus, 1985), and shows that groups can be less-than-optimal users of information. When Stasser and Titus (1985) asked groups of 4 members to reach a judgment about a hypothetical candidate for a student body president, they found that groups often focused discussion on shared rather than unshared information, and as such, the groups produced biased judgments about the potential candidate. In a replication of Stasser and Titus (1985), Wittenbaum and Stasser (1996) showed that groups sometimes fail to uncover new information due to the dominant role of shared information during group deliberation. Wittenbaum and Stasser (1996) noted some potential mechanisms driving the effect, most of which were social, e.g., a group member who discussed shared information increased their perceived credibility.

Other social aspects contribute to group deliberation and, ultimately, collaborative performance levels. Some empirical evidence of the social factors contributing to collaborative decision-making suggests groups sometimes fall short of the predictions of statistical models of collaborative decision making because social influences bias individual group members' judgments (Kerr & Tindale, 2004; Davis, 1992; Lorenz, Rauhut, Schweitzer, & Helbeing, 2011). In an experiment reported by Lorenz et al. (2011), for instance, 12-member groups estimated crime statistics for various geographical locations. Each group member made five estimations for each location, and provided a confidence estimate on a Likert scale ranging from 1 (very uncertain) to 6 (very certain) for their first and last estimates.

All participants completed three information conditions (Lorenz et al., 2011). In the "aggregated information condition", group members' estimates for each round after the first were based on the mean of group members' estimates of the immediately previous round (i.e., only one round of estimation aggregated to produce a mean). In the "full information" condition, group members could base estimations on a figure depicting the trajectories of all group members' estimates from all previous rounds of that specific estimation. The "no information" condition operated as a control condition in which group members' estimates were unavailable. The results indicated that the aggregated information and full information conditions reduced the diversity of group members' estimates without raising judgmental accuracy. In other words, group members' estimates converged but did not improve, a pattern referred to by Lorenz et al. (2011) as a *social influence effect*. The authors argued that the observed social influence effect reflected groups' inability to make use of information exchange. The aggregated and full information conditions also showed a *range reduction effect* where group members based estimates on predictions that were narrowly distributed around the wrong value, limiting the potential benefit of collaboration. A third and final detrimental effect, *the confidence effect*, was found to undermine the wisdom of crowds effect by boosting individuals' confidence in estimates without an associated increase in collaborative accuracy.

Another important social factor contributing to collaborative performance is motivation. Multiple reviews of collaborative performance (e.g., Davis, 1990; Karau & Williams, 1993; Kerr & Tindale, 2004; Levine & Moreland, 1990) note that collaborating team members might put forth less effort when working in the presence of group members, compared to when working independently (i.e., social loafing; Karau & Williams, 1993; Kerr & Tindale, 2004; Latané et al., 1979), contributing to group motivational losses. Alternatively, group members can also put forth more effort when working in the presence of other group members (i.e., social facilitation; Forsyth, 1998; Zajonc, 1965), showing group motivational gains (Kerr & Tindale, 2004; Latané, Williams, & Harkins, 1979).

A number of mechanisms contribute to groups' motivational losses (Kerr & Tindale, 2004; Latané et al., 1979; Steiner, 1972). Smith, Kerr, Markus, and Stasson (2001) note that some group members might put forth less effort than others because there is an opportunity to 'free ride' on the efforts of other group members, or an unwillingness to do the work that other group members could be doing.

Karau and Williams' (1993) *collective effort model* (CEM) argues that social loafing reflects the contingency between a group member's effort and that group member's valued outcomes. As such, the CEM predicts that group members will put forth effort when performing a collaborative task to the extent that their efforts are perceived as highly instrumental in obtaining valued outcomes. Conversely, they will put forth less effort if their contributions are perceived as not linked to collaborative performance (low instrumentality) or if the potential collaborative outcomes are not valued. Group members' expectancy, the degree to which each group member believes high levels of effort correspond with high collaborative performance levels, also play a role in the CEM. The authors argue that individual group members' efforts can be conceptualised as the product of group members' perceived instrumentality, expectancy, and the outcome values associated with a particular collaborative task.

Mechanisms underlying group motivational gains are also of interest to researchers (Karau & Williams, 1997; Kerr & Tindale, 2004). Social facilitation theory (Forsyth, 1998; Zajonc, 1965) suggests individual group members put forth more effort when collaborating purely due to the mere presence of other group members. Another proposal, social compensation theory (Williams & Karau, 1991), argues that individual group members might increase their efforts when participating in collaborative tasks to compensate for the assumed poor performance of other group members. The poorest performing group members might increase their efforts in the event that they are aware of the discrepancy between their and other group members' performance levels, a phenomenon termed the Köhler effect (Köhler, 1926; Witte, 1989). Witte (1989) noted that the Köhler effect is most likely to occur when the poorest performing group members define a group's performance level. Finally, Erev, Bornstein, and Galili (1993) demonstrated that the potential for intragroup competition might also increase individual group members' motivation levels.

Non-collocated groups. Social factors also contribute to the value of collaboration when group members are not collocated (Chidambaram & Tung, 2005; Kiesler & Cummings, 2002). Sometimes referred to as virtual, distributed, or remote teams, group members are required to collaborate from different locations and communicate electronically (Kiesler & Cummings, 2002; Moon & Sproull, 2001; Sproull & Kiesler, 1991). Kiesler and Cummings (2002) noted that distributed teams are likely to suffer from social as well as physical distance. This social distance encompasses a number of social factors, such as decreased group cohesion or group identity, limited or no face-to-face communication, and increased social diversity of teams, that potentially reduces collaborative performance (Kiesler & Cummings, 2002).

The reduction of face-to-face communication in remote teams is particularly interesting because of the empirical evidence showing its powerful effects on collaboration (Deutsch, 1958; Kerr & Kaufman-Gilliland, 1994; Kiesler & Cummings, 2002). Kerr and Kaufman-Gilliland (1994) asked participants to engage in an investment game in which some teams were collocated and others were not. Working in groups, some group members were given 5 minutes to discuss the game, while other group members did not communicate. Face-to-face communicating group members were far more likely to cooperate with the group than were non-face-to-face communicating group members and, even more interesting, this effect was not evident when group members could hear the face-to-face communication but not participate in it. Groups with more cooperative group members outperformed groups with less cooperative group members.

But despite limited face-to-face communication, remote teams can produce collaborative performance levels on par with collocated teams (Purvanova, 2014; Scott & Wildman, 2015). Shachaf (2008) interviewed 41 team members of distributed teams from nine countries employed by Fortune 500 corporations in attempt to better understand how virtual team environments influence the effectiveness of collaboration. The interviews focused on the effects of cultural diversity and information and communication technology (ICT) on virtual team effectiveness. Responses indicated that cultural diversity influenced collaborative decision-making positively but also influenced communication negatively. The detrimental effect of intercultural communication was moderated, though, by effective ICT.

More recently, Purvanova (2014) explored the discrepancy between the findings of experimental literature and field research on virtual teams. Purvanova (2014) noted that multiple meta-analyses of the experimental literature on virtual teams (e.g., Baltes, Dickson, Sherman, Bauer, & LaGanke, 2002; Benbasat & Lim, 1993; Rains, 2005) portray distributed groups negatively, inferior to traditional face-to-face teams, whereas field studies tend to show positive outcomes of remote teaming. One potential explanation for the discrepancy, according to Puranova's (2014) qualitative review of experimental and field-based literatures, is the lack of ecological validity in experimental studies. More specifically, Purvanova (2014) argued that the reality of virtual teams is not accurately replicated in common experimental tasks, which require group members to work for short periods of time on inconsequential tasks. Ultimately, Purvanova (2014) found virtual teams a viable alternative to traditional, face-to-face teams, and noted the importance of experimental research to better understand collaborative processes and the outcomes of ad hoc teams, groups created to address a specific issue within a restricted timeframe.

Current aims

Research on collaborative visual search is important to better understand the conditions under which teamwork produces more efficient search than independent search. This thesis investigated joint search in a naturalistic signal detection task. A series of experiments across two studies compared the collaborative sensitivity of two-person teams to that of individuals and the predictions of a uniform weighting information integration model and explored collaborative search efficiency with little target information.

Broadly, this thesis aimed to better understand collaborative visual search. More specifically, this thesis had two main aims. First, to replicate previous findings (i.e., collaborative search performance the same or better than the predictions of the uniform weighting model of information integration) and extend them to a signal detection task using naturalistic stimuli. Second, to investigate a joint search strategy when little target information is available to observers.

CHAPTER 2:

STUDY 1

The following manuscript entitled, *Collaborative Search in a Mock Baggage Screening Task: A Bayesian Hierarchical Analysis*, is currently available on PsyArxiv Preprints (https://doi.org/10.17605/OSF.IO/8975X). The version of the manuscript presented here is a revised version.

Both authors were involved in the formulation of the study concept and design, and data analysis. Ali Enright collected the data and completed the initial draft of the manuscript. Jason McCarley edited multiple revisions of the manuscript.

Abstract

Signal detection theory provides models of information integration that allow researchers to predict and benchmark collaborative performance in a visual search task. Naturalistic stimuli, however, may not conform to the simplifying assumptions—specifically, assumptions of equal-variance signal and noise distributions and stochastically independent observers—that are often made to make collaborative signal detection models tractable. Here, we used Bayesian hierarchical modeling of receiver operating characteristics to circumvent this difficulty. Participants (*N* = 28–32 per experiment) performed a simulated baggage x-ray screening task, working alone or in teams of two. Team performance was compared to the predictions of two versions of a uniform weighting model of information integration, one that assumed stochastically independent judgments from the two members of a team and one that allowed for correlated judgments. Across four experiments, teams fell short of the uncorrelated-judgment model's predictions, but outperformed predictions based on the observed correlations in individual judgments. Results imply motivational effects that improve individual searchers' effort under collaborative conditions, or collaborative strategies that effectively decorrelate the individual searchers' judgments.

Introduction

Visual search in domains such as transportation security screening and medical image reading is often collaborative, requiring multiple searchers to reach a joint decision. Signal detection theory (SDT; Green & Swets, 1966; Macmillan & Creelman, 2005), a model of how decision makers reach discrete judgments from probabilistic evidence distributions, provides a framework for predicting and assessing collaborative efficiency. A binary signal detection task requires a decision maker to judge whether a stimulus was drawn from a distribution containing only noise or containing noise with an embedded signal (Green & Swets, 1966; Macmillan & Creelman, 2005). The decision maker reaches a judgment by reducing the evidence for or against the presence of a signal to a scalar decision variable and comparing its value to a criterion. A positive judgment results when the decision variable exceeds the criterion, and a negative judgment results otherwise. On a signal-plus-noise trial, a correct response is a *hit* and an incorrect response is a *miss*. An accurate response on a noise trial is a *correct rejection*, and an inaccurate response on a noise trial is a *false alarm*. The ability to accurately distinguish signal-plus-noise from noise-alone distributions is *sensitivity*, while the decision-maker's general tendency to respond 'yes' or 'no', as determined by the criterion, is *bias*. The standard form of the SDT model, the equal-variance Gaussian model, assumes that signal and noise distributions of decision variables are normal with different means but equal variances. Within this model, sensitivity is conventionally measured by d'(Green & Swets, 1966), the distance between the means of the signal and noise distributions, in units of the standard deviation.

Collaborative sensitivity

The level of sensitivity achieved by a collaborative team is a function of the individual team members' sensitivity levels, their information pooling strategy, and the correlation between their judgments (Bahrami et al., 2010; Sorkin & Dai, 1994; Sorkin, Hays, & West, 2001). Ideal collaborative performance comes from the *Optimal Weighting Model* (OW; Bahrami et al., 2010; Sorkin & Dai, 1994; Sorkin et al., 2001). The OW model assumes that individual team members' judgments are averaged in a weighted manner to produce a team decision. The team's optimal decision variable, X_{team}, is given by,

$$X_{team} = \sum X_i d'_i$$

where X_i is the decision variable for team member *i*, and d'_i is team member *i*'s sensitivity. Assuming that the team members' judgments are stochastically independent, team sensitivity under the OW model is,

$$d'_{OW} = \sqrt{\sum d'^2_i}. \qquad [1]$$

The Uniform Weighting Model (UW; Sorkin & Dai, 1994) is similar to the OW model, but assumes that individual team members' signal likelihood judgements are weighted equally when they are averaged to reach a group judgement. Again, assuming an equal-variance Gaussian model, group d' under the UW model is,

$$d'_{UW} = \frac{\sum d'_i}{\sqrt{N}} \qquad [2]$$

Bahrami et al. (2010) note that uniform weighting is equivalent to a circumstance in which team members do not directly communicate signal likelihood estimates, but instead convert an internal evidence representation into a confidence rating through a standard normal transformation. When collaborators are equally sensitive, the UW model produces the same predictions as the OW model. Otherwise, the unweighted model produces lower sensitivity.

Collaborative performance in at least some perceptual and statistical decision making tasks approaches levels predicted by the UW model (Bahrami et al., 2010; Sorkin et al., 2001). Bahrami et al. (2010) presented dyads with two stimulus intervals each containing six Gabor patches. A single target, a patch with slightly elevated contrast, was present in either the first or second interval, and the participants' task was to report the target interval. Dyad members first provided individual responses, without consulting each other, and in the event their responses conflicted, provided a joint response. In Experiment 1, stimuli were approximately equally discriminable for all participants. In Experiment 2, visual noise was added to the stimuli for one participant within each dyad in order to elicit different levels of sensitivity for the two dyad members. Results from both experiments were consistent with the UW model. Bahrami et al. (2012) extended these findings to conditions of nonverbal communication between team members.

Sorkin, Hays, and West (2001) also assessed collaborative decision making performance. Participants performed a signal detection task individually and in groups ranging in size from 2–7 members. On each trial, participants were presented with a display of nine analog gauges. Values of the gauges were drawn from one of two equal-variance Gaussian distributions, one of higher mean than the other, and the participants' task was to decide which distribution the values came from. Groups performed better than individuals, but with efficiency lower than predicted either by the OW model or by a simple majority rule model. Deviations from optimal group performance did not seem to result from inefficient combination of the individual participants' decisions. Rather, group performance was poorer than expected because individual observers put less effort into the task as group size increased (shown by decomposing a measure of group efficiency into two components – group member effort and group decision aggregation efficiency – and comparing both components' contribution to the overall group detection performance).

Another study of collaborative visual search likewise found what appeared to be highly efficient collaboration. Malcolmson, Reynolds, and Smilek (2007) did not directly compare individual to team performance but compared the collaborative sensitivity levels of empirical and nominal teams. Empirical teams performed a visual search task sitting together at the same computer. Nominal teams were formed by combining the judgments of two members of two team members who performed the task in isolation, viewing a sequence of stimuli. Nominal teams thus provided a benchmark of the performance that would be expected from a UW model in which team members contributed judgments without interacting, and comparisons to the empirical teams revealed the costs or benefits to performance that might arise from active collaboration between team members. Empirical teams outperformed nominal teams, implying performance better than expected from a strict UW strategy.

Participants reported a strategy of dividing the search display into halves, with each team member taking responsibility for searching one half.

Generalizing the collaborative models described above to naturalistic contexts is challenging, however, for at least two reasons. First, team members performing a collaborative task outside the lab are likely to render correlated judgments. As noted, the predictions in Equations 1 and 2 assume that individual team members contribute stochastically independent judgments, which are then summed to produce a team decision variable. Correlations between team members' responses reduce the benefit of collaboration. In the extreme, when collaborators produce identical responses every trial, collaboration produces no benefit at all.

The dependence between collaborators' responses is determined by the relative strength of unique and common noise in their information encoding (Sorkin et al., 2001), where unique noise is the variance isolated to an individual team member and common noise is that shared among team members. For example, consider a pair of transportation security officers jointly inspecting a baggage x-ray. Unique noise might arise from differences in the searchers' internal sensory noise, or from differences in their search patterns. Common noise might arise from ambiguities or degradations in the x-ray image itself, from variations in target salience (Mello-Thoms, 2006), or from similarities in the viewers' search patterns. Many past studies of collaboration in signal detection tasks have ensured uncorrelated judgments by presenting observers stimuli generated separately and independently (e.g., Bahrami et al., 2010, 2012; Sorkin et al., 2001). In naturalistic tasks, however, as in the case of the transportation security officers' joint inspection above, collaborators are likely to view the very same stimulus, introducing a source of common variance. Unfortunately, because the decision variables that are the basis of the observers' judgements are internal and unobservable, the degree to which they are correlated in such cases is difficult to know (Metz & Shen, 1992).

Naturalistic tasks are also likely to violate the assumption of equal-variance signal and noise distributions on which the formulas above are based; because the signal-plus-noise evidence distribution contains the variance associated with the signal as well as that associated with the noise, its standard deviation will exceed that of the noise distribution anytime the properties of the signal are not deterministic (Swets, 1986).

Bayesian hierarchical analysis of the ROC

Analysis of the receiver operating characteristic (ROC; Macmillan & Creelman, 2005; Morey, Pratte, & Rouder, 2008; Swets, Tanner, & Birdsall, 1961) provides a method for circumventing both the problem of stochastic dependence between observers and the problem of unequal signal and noise variance. A ROC plots hit rate in a signal detection task as a function of the false alarm rate, holding sensitivity constant. Empirically, the multiple hit and false alarm rate pairs needed to plot the ROC are collected by varying the decision maker's response criterion while holding task difficulty fixed. This can be done by manipulating signal rate or response payoffs across blocks of trials, or by collecting confidence-rated judgments in a single block of trials (Macmillan & Creelman, 2005).

Plotting the ROC in standardized normal space produces a *z*ROC, which can be used to test the assumptions of equal variance Gaussian evidence distributions. Assume, following convention, that the evidence distribution for noise trials is Gaussian with mean 0 and standard deviation 1, and the signal plus noise distribution is Gaussian with mean μ_S and standard deviation σ_S. Then, the *z*ROC is a linear function,

$$z(H) = \frac{\mu_s}{\sigma_s} + \frac{1}{\sigma_s} \times z(F)$$

where H and F are the raw hit and false alarm rates, respectively, and z is the inverse normal transformation (Wickens, 2002). Normal evidence distributions, that is, imply a linear *z*ROC with a slope equal to the inverse of the standard deviation of the signal-plus-noise evidence distribution, and a *y*-intercept equal to the mean of the signal-plus-noise distribution divided by the standard deviation of the signal-plus-noise evidence distribution. A useful sensitivity measure derived from the zROC is the distance measure d'_e (Macmillan & Creelman, 2005),

$$d'_e = \frac{2\mu_s}{1+\sigma_s}.$$

When noise-alone and signal-plus-noise distributions are equal-variance, d'_e reduces to d'. Otherwise, d'_e may exceed or fall below d'(Macmillan & Creelman, 2005).

Notably, the UW model can be adapted to predict collaborative signal detection performance in ROC space without the assumption of equal-variance distributions. As noted, the model assumes that group judgments are based on the unweighted average of the individual team members' judgments. The sum of two normally-distributed independent random variables X_1 and X_2 is itself normally distributed, with mean and variance, respectively, equal to the summed means and variances of X_1 and X_2 (Macmillan & Creelman, 2005). That is, if

$$X_1 \sim N(\mu_1, \sigma_1{}^2)$$

$$X_2 \sim N(\mu_2, \sigma_2{}^2)$$

then,

$$X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1{}^2 + \sigma_2{}^2).$$

Assume that the noise distribution for each individual team member is normally distributed with mean 0 and standard deviation of 1, and that signal-plus-noise distributions for the two group members are normally distributed with means μ_1, μ_2 and standard deviations σ_1, σ_2. Then, assuming uncorrelated judgments from the two team members, the noise distribution for the group decision variable is,

$$X_{n_1+n_2} \sim N(0,2\ )$$

and the signal-plus-noise distribution for the group decision variable is,

$$X_{s_1+s_2} \sim N(\mu_1 +\ \mu_2, \sigma_1^2 + \sigma_2^2\ ).$$

These values specify the predicted ROC for a two-person team based on a UW model assuming stochastically independent observers. We will call this the $UW_{\rho\ =\ 0}$ model.

The same approach can be adapted to predict the UW ROC for stochastically dependent team members. However, this requires that the correlation between the members' judgments is known (Metz & Shen, 1992). An alternative approach for modelling team performance from judgments is to calculate a mock team decision variable from individual observers' individual judgments of yoked stimuli

(Malcolmson et al., 2007; Metz & Shen, 1992; Sorkin et al., 2001). Averaging trial-by-trial confidence ratings from paired observers in an unweighted manner provides an estimate of the responses that would be produced by a UW strategy, and inherently incorporates stochastic dependency between the participants' individual judgments. We will call responses generated this way *mock UW team* judgments. These responses can then be analysed in just the same way as observed responses. As noted above, Malcolmson et al. (2007) found evidence for higher sensitivity in empirical teams than in nominal teams, implying performance better than expected from a simple UW strategy.

Conventional analyses of ROCs for multiple observers aggregate results over stimulus items, observers, or both. As Morey, Pratte, and Rouder (2008; Pratte & Rouder, 2012; Pratte, Rouder, & Morey, 2010) have shown, though, this practice can distort estimates of signal-distribution variance and *z*ROC curvature. As an alternative method of analyzing group ROC data, Morey et al. (2008) recommend a Bayesian hierarchical approach (cf. Rouder & Lu, 2005). Here, we use Bayesian hierarchical ROC analysis to examine collaborative visual search using naturalistic stimuli, simulated baggage x-rays. Participants searched baggage x-rays for hidden knives, working either individually or in teams of two. Confidence ratings were collected to allow analysis of the ROC and observed team performance was compared to the performance of the mock UW teams and the predictions of the $\mathrm{UW}_{\rho=0}$ model.

Experiment 1

In Experiment 1, participants performed the simulated baggage screening task individually, in separate testing rooms, or collaboratively, sitting together in the same room.

Method

Participants

Sixteen pairs of undergraduate students (25 female, M_{age} = 24.3, *SD* = 8.0) were recruited via the School of Psychology research participant pool at Flinders University. Each participant received $20AU in exchange for participation. All participants demonstrated normal color vision (determined using Ishihara test) and normal or corrected-to-normal visual acuity (tested using a standard visual acuity chart in the lab).

Apparatus and Stimuli

Stimuli for the visual search task were presented on a 370 mm x 300 mm Samsung monitor (model S24D590PL), with a resolution of 1920 x 1080 pixels and a refresh rate of 85 Hz. Stimulus display and response collection were controlled by software custom written in PsychoPy (Peirce, 2007, 2009). Displays were viewed from a distance of roughly 570 mm, though viewing distance was not constrained.

Stimuli, the same used in an earlier study (McCarley, 2009), were color x-ray images of various bags (e.g., backpacks, suitcases, briefcases) containing everyday items (e.g., hair dryers, keys, clothes, portable electronics). Stimuli were created by combining images of individual objects using Photoshop (Adobe Systems Inc., San Jose, CA, USA) and MATLAB (The Mathworks, Natick, MA, USA). Images were combined through multiplicative blending, producing the appearance of transparency. Stimulus size ranged from 9.55° x 6.61° to 23.32° x 21.88°. A pool of 600 images was created for use as target-absent stimuli. Target-present stimuli were created by randomly choosing target-absent stimuli with replacement from the pool of 600, randomly rotating them by 0°, 90°, 180°, and 270°, and then inserting knives at random locations and random orientations of 0°, 45°, 90°, 135°, 180°, 225°, 270° or 315° in the picture plane. Targets were inserted into the target-absent images using multiplicative blending, producing the appearance of transparency. Five knives, ranging in size from roughly 2.5° x 0.3° to 4.0° x 0.5°, served as targets**.** Only one knife was present in each target-present stimulus image. Figure 2.1 presents a sample stimulus.

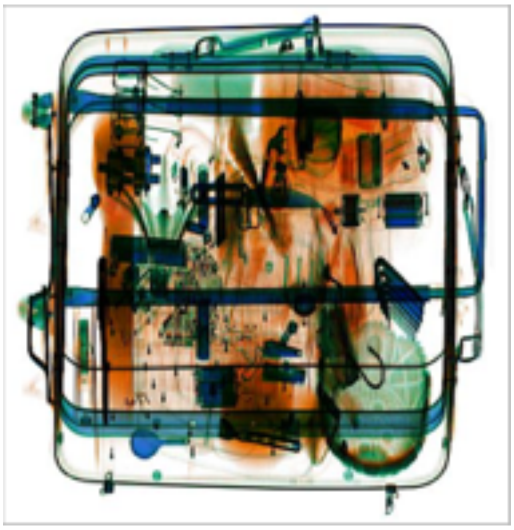

Figure 2.1. An example of a target-present baggage x-ray.

Stimulus images were randomly divided into two sets, A and B, both of which contained 100 target-present and 150 target-absent images. (Target-present and –absent trials were unbalanced in order to encourage a conservative response bias as in earlier work; McCarley, 2009). Odd-numbered teams used set A for individual conditions and set B for collaborative conditions, and even-numbered used the reverse assignment.

Procedure

Participants completed the visual search task both individually, in separate testing rooms, and collaboratively, sharing one computer in the same testing room. Instructions were presented onscreen at the start of the experimental session. The instructions asked participants to imagine that they were transportation security screeners in an airport and explained that their task was to decide if a knife was present (signal-plus-noise event) or not (noise-alone event) in each x-ray image. Images of the five target knives were presented onscreen below the instructional text.

The search task began after participants had read and affirmed that they understood the instructions. Each trial began with a fixation interval lasting 1000ms, after which the stimulus image was presented for free viewing, with a rating scale below it. Responses were made via mouse click on the six-point scale, which included the options *Definitely Yes, Probably Yes, Guess Yes, Guess No,* *Probably No, Definitely No*. *Definitely Yes, Probably Yes,* and *Guess Yes* were treated as correct responses for target-present trials, whereas *Guess No, Probably No,* and *Definitely No* were treated as correct responses for target-absent trials. A feedback message of ‘You found a threat!’, ‘Good judgement’, ‘You missed a threat!’, or ‘False alarm’ followed a hit, correct rejection, miss, or false alarm, respectively.

Participants completed one block of 250 collaborative trials and one block of 250 individual trials. The order of blocks and choice of testing room for the collaborative condition were counterbalanced across teams. Trial order was randomized within blocks and yoked across participants in the individual search conditions in order to ensure that any potential influence of stimulus order on performance was matched across team members.

Analyses

Data were analyzed in RStudio (www.rstudio.com) using the Hierarchical Bayesian Analysis of Recognition Memory package (HBMEM; Morey, Pratte, & Rouder, 2008; Pratte, Rouder, & Morey 2009; Pratte & Rouder, 2012), which contains functions for fitting hierarchical versions of equal and unequal variance Gaussian signal detection models to confidence rating data. The model assumes an additive effect of observer on the mean separation between noise-alone and signal-plus-noise distributions, and where variance is not held fixed across observers, an additive effect of observer on the log variance of the signal-plus-noise distribution (Pratte & Rouder, 2012). Model-fitting functions employ Bayesian Markov chain Monte Carlo (MCMC) sampling procedure, assuming diffuse priors on model parameters. The model was run for 10,000 burn-in iterations, followed by 50,000 iterations for analysis.

Three versions of the model were fit. The first (EV) assumed equal variance noise-alone and signal-plus-noise distributions. The second (UV, fixed S^2) allowed the variance of the signal-plus-noise distribution to differ from that of the noise-alone distribution but assumed that it was fixed across observers (Pratte et al., 2009). The third (UV, free S^2) also allowed unequal variance between the noise-alone and the signal-plus-noise distributions, but with the variance of the signal-plus-noise distribution free to vary across observers (Pratte & Rouder, 2010). Models within the HBMEM package also allow for estimates of item-level stimulus effects. This requires, however, that individual stimulus items be crossed with signal/noise condition. Here, individual stimulus images were nested within signal and noise conditions. Model fits were compared using the deviance information criterion (DIC), a statistic that measures the quality of model fit, incorporating a penalty for the number of functional model parameters (Spiegelhalter, Best, Carlin, & van der Linde, 2002). Smaller values indicate better performance.

Because two individual participants were associated with each team, the experimental design did not lend itself to a paired-samples analysis of the collaborative versus team conditions. Search condition was therefore treated as a between-subject manipulation. Before fits of the three models were generated, data were submitted to a preliminary run of the UV, free S^2 model (1000 burn-in iterations, 10000 iterations for analysis) to identify any individuals or teams who failed to meet an inclusion criterion of $d'_e \geq 0.5$. If any individual participant or team failed to meet the minimum d'_e for inclusion, all data for that team (both single observer and collaborative data) were excluded from further analysis. A total of four teams were excluded (one team in Experiment 2, one team in Experiment 3, and two teams in Experiment 4).

Predictions for the $\mathrm{UW}_{\rho=0}$ model were calculated from the hierarchical group mean parameter estimates of μ_{n}, μ_{s}, and σ_{s} at each iteration of the MCMC process. Note that because the hierarchical analysis produced only a single group-level distribution for the individual search condition (i.e., separate estimates were not made for each observer within a team), model predictions at this level assumed that the two searchers within a team were equally sensitive. This reduces the OW model to the UW model. Mock UW team predictions were generated by averaging the two participants' responses for each trial of the individual search condition, rounding the result to put the responses on a 6-point scale, then submitting the data to the HBMEM model. Averaging team member responses in this way inherently includes the correlation between team members' judgments, alleviating the need to know the exact correlations between team members' decisions.

Data reported in the text below are the means and 95% Bayesian credible intervals (BCI) of the posterior distributions produced by the model. Plots of data were generated in R using the *ggplot2* package v 2.2.2 (Wickham & Chang, 2016), including the geom_density function for plots of posterior distributions.

Results

All participants and teams in Experiment 1 met the minimum d'_escore for inclusion. Table 2.1 shows the DIC values for all three variants of the model, for Experiments 1-4. In Experiment 1, the UV, fixed S^2 produced the lowest DIC value. However, the UV, free S^2 model produced the lowest DIC values for the remaining experiments. For consistency, the results of UV, free S^2 model are reported for all four experiments, though comparison of the UV fixed and free S^2 models suggested no substantial differences in the results for any of the experiments.

Figure 2.2 shows the post burn-in MCMC chains for the model fitting procedure. By inspection, chains appear to have converged.

Table 2.1. DIC values for the EVSD, UV, S^2 fixed and UV, S^2 free for Experiments 1-4

	DIC values		
	EV	UV, S^2 fixed	UV, S^2 free
Experiment 1	33511.96	32438.76	32446.21
Experiment 2	31432.09	31430.43	31426.83
Experiment 3	32462.82	32464.74	32460.56
Experiment 4	32478.86	32476.45	32471.50

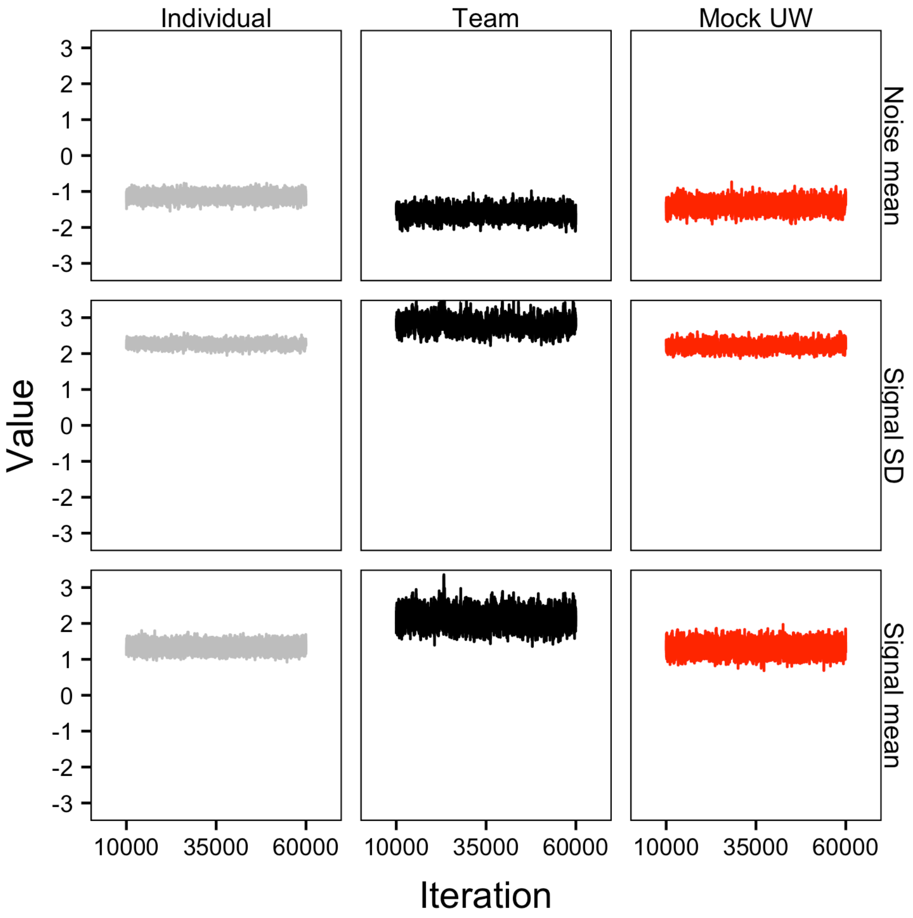

Figure 2.2. MCMC chains for Experiment 1. Columns represent task condition, rows represent estimated parameters.

Figure 2.3 shows the *z*ROCs for the single observers, teams, the UW$_{\rho=0}$ model, and mock UW teams, based on estimates of the group-level parameters. The *z*-slopes for the signal and noise distributions were less than 1.0 ($M = 0.53$ for single observers, and $M = 0.40$ for teams) indicating that the signal distribution had a larger variance than the noise distribution, as expected.

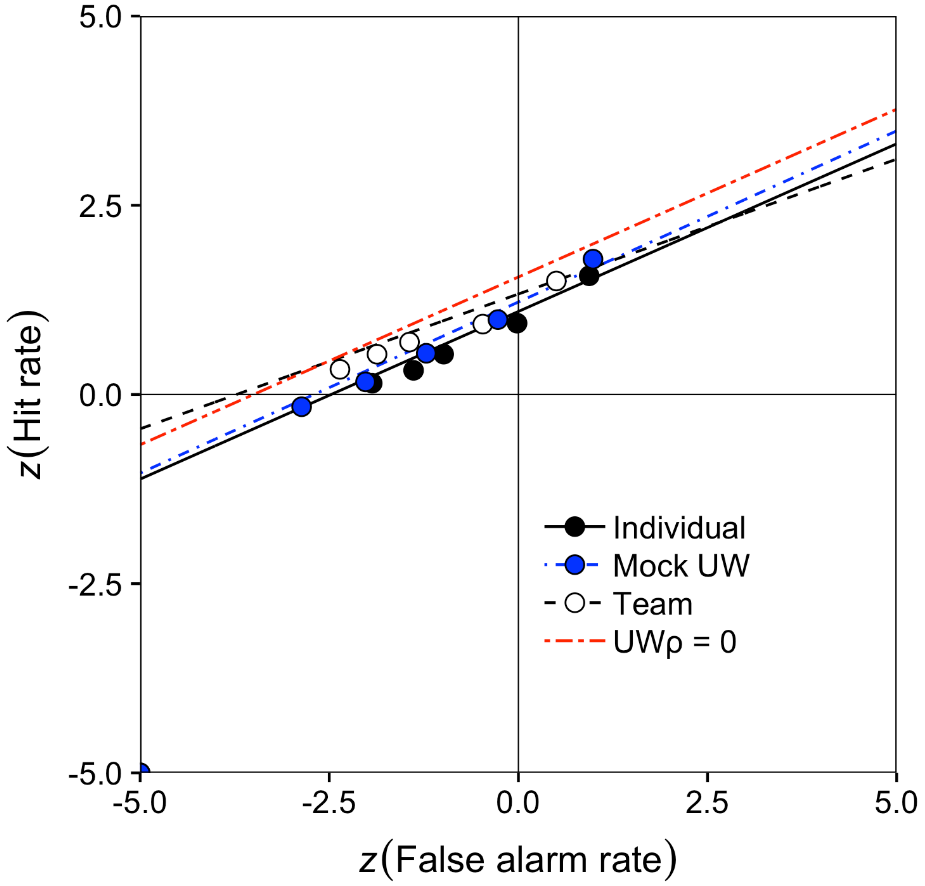

Figure 2.3. *z*ROCs for Experiment 1, with group mean data points superimposed.

Figure 2.4 shows the estimated posterior distribution of d'_e scores, again based on estimates of the group-level parameters. Figure 2.5 presents the distributions and 95% BCIs of the difference scores between observed and model-predicted team performance levels. Teams (M = 1.96, BCI[1.75, 2.17]) outperformed the single observers (M = 1.52, BCI[1.36, 1.68]), mean difference = 0.44, BCI[0.18, 0.70]. Mean team sensitivity fell roughly midway between that of the mock UW teams (M = 1.68, BCI[1.46, 1.89]) and the $\mathrm{UW}_{\rho=0}$ model (M = 2.15, BCI[1.93, 2.37]). Error scores did not differ credibly from zero for either model.

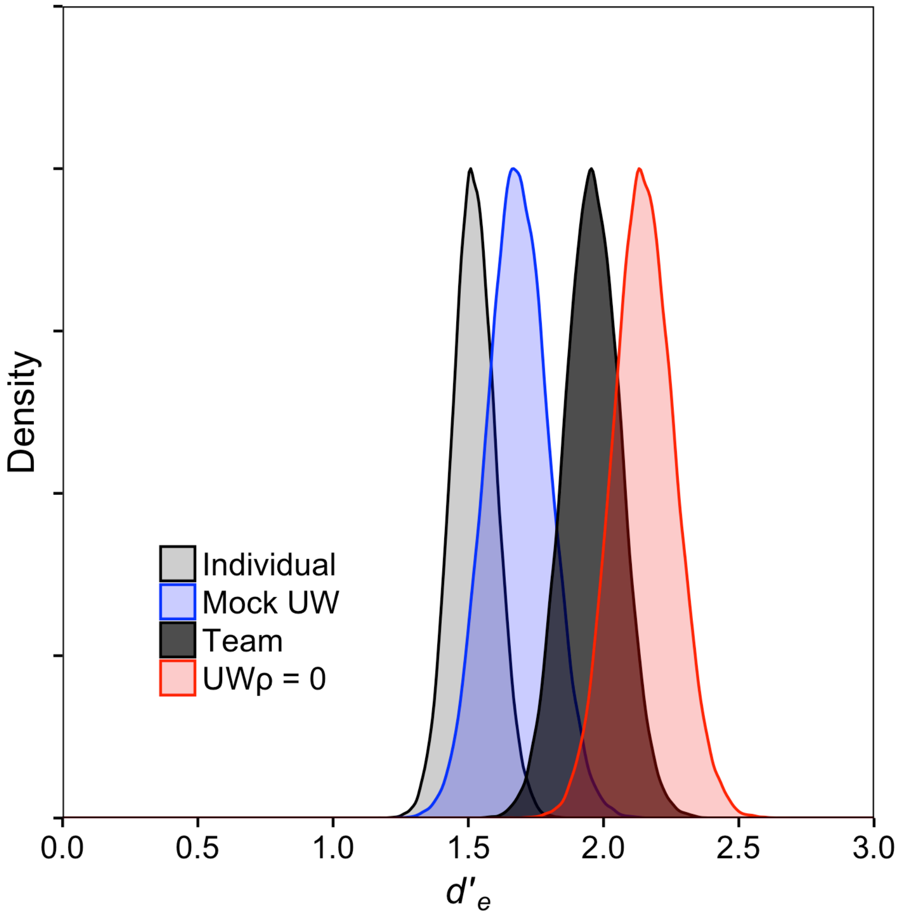

Figure 2.4. Posterior distributions of d'_e for single observers (light gray), teams (dark gray), the UW$_{\rho=0}$ model (blue), and mock UW teams (red) in Experiment 1.

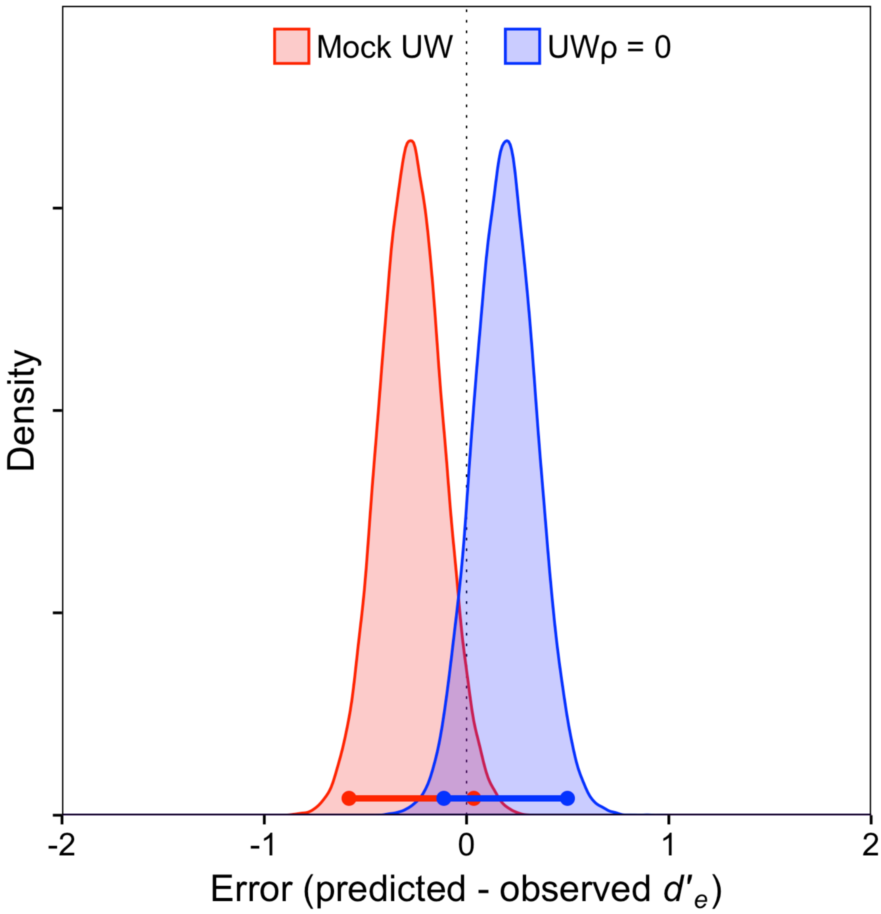

Figure 2.5. Difference scores between observed team performance and $UW_{\rho=0}$ model (blue) and mock UW team (red) performance in Experiment 1. Error bars are 95% credible intervals on the difference between observed and predicted scores.

An additional analysis compared observed scores to the performance of the mock UW teams and the $UW_{\rho=0}$ model at the team-by-team level. Figure 2.6 shows the difference between observed and predicted team performance for all 16 teams. The mock UW model underestimated sensitivity for four (teams 2, 14, 15, and 16) predictions differed credibly from the zero-error point of observed team performance. The $UW_{\rho=0}$ model overestimated sensitivity for one team (team 13) and underestimated it for two others (teams 14 and 16).

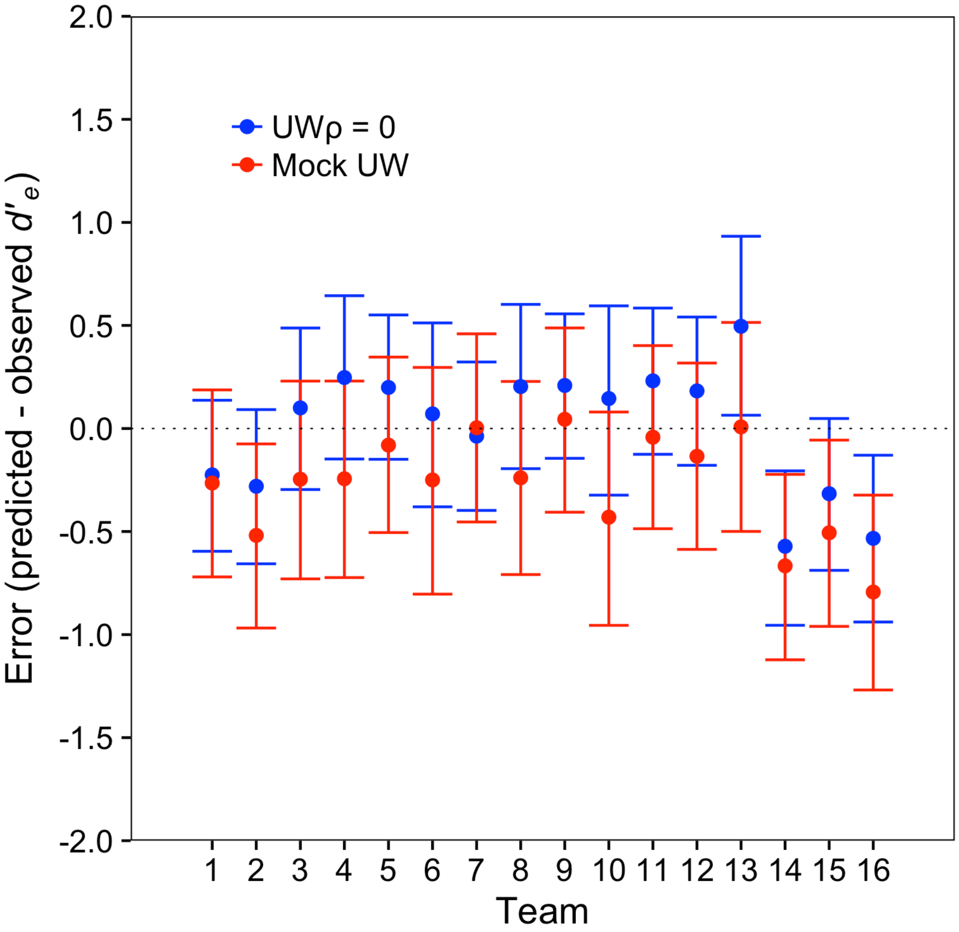

Figure 2.6. Difference scores between observed team *d'$_e$* and UW$_{\rho=0}$ model (blue) and mock UW team (red) predictions in Experiment 1. Error bars are 95% credible intervals on difference between observed and predicted scores.

Discussion

Observed team performance levels fell between the predictions of the correlated and uncorrelated UW models. Although past work has found collaborative visual search performance consistent with the uncorrelated UW model, the finding that performance here fell short of this level was expected; collaborating participants both viewed the same image each trial, creating a significant source of shared variance in their judgments. This shared variance should have engendered a strong stochastic dependence in the paired observers' judgments, pushing performance below the predictions of the uncorrelated UW model. What's more surprising is the finding that team performance trended higher than the predictions of a UW model that accounted for the correlations in individual team members' judgments. In other words, although the effect fell just short of being credible at the 95% level, teams trended toward performance better than expected given their correlated responses.

A potential explanation for this unexpected result is that estimates of collaborative efficiency were biased upwards by social processes unrelated to information integration strategy. The UW model assumes that team members' sensitivity levels are the same under individual and collaborative conditions, and that the benefits of collaboration arise strictly from the combination of team members' judgments. It is possible, though, that participants were more motivated or put forth more effort when working side-by-side in the same testing room than when working alone in separate rooms (Olsen, Bahrami, Roepstorff, & Frith, 2010), achieving higher levels of sensitivity in their individual judgments even before combining judgments to reach a team decision. This suggests that collaborative searchers might not outperform the mock UW model if they perform the individual search task in the same room. Notably, control experiments by Malcolmson et al. (2007) tested this possibility in a simpler form of visual search task and found no evidence that simply working in the same room as another participant produced benefits on par with those of active collaboration. Nonetheless, to rule out this potential confound, Experiment 2 replicated the design of the first experiment, but asked the participants to complete both the individual and collaborative conditions in the same testing room.

Experiment 2

Participants

Sixteen pairs of undergraduate students (22 female, M_{age} = 22.3 *SD* = 7.4) were recruited via the School of Psychology participant pool at Flinders University. Each participant received $20AU in exchange for participation. All participants demonstrated normal color vision and normal or corrected-to-normal visual acuity.

All stimuli and procedures were the same as Experiment 1 except that participants completed the individual condition in separate workstations within the same testing room. In the individual condition, one participant faced a display located on the north wall, while the other faced a display located on the east wall of the same room. In the individual testing conditions, participants were instructed to look strictly at their own display and to refrain from communicating. When one participant completed the individual condition, he or she exited the testing room and waited in an empty room for the other member to finish.

Results

Data from one team were excluded because one team member ($d'_e = 0.18$) failed to meet the inclusion criterion, leaving data from 15 teams for analysis. Figure 2.7 presents the post burn-in MCMC chains for the model-fitting procedure. By inspection, chains appear to have converged.

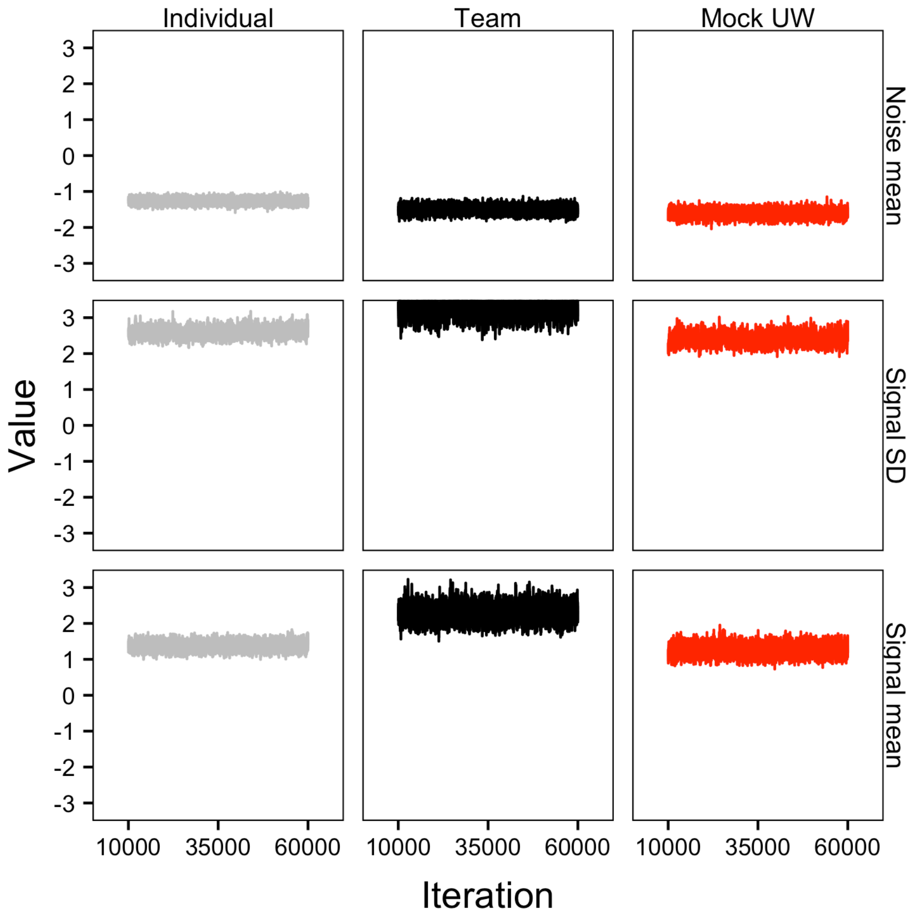

Figure 2.7. MCMC chains for Experiment 2. Columns represent task condition, rows represent estimated parameters.

Figure 2.8 shows the *z*ROCs for the single observers, teams, the $UW_{\rho=0}$ model, and mock UW teams, again based on estimates of the population-level parameters. The *z*-slopes for the signal and noise distributions were similar to those of Experiment 1 (*M* = 0.60 for single observers, and *M* = 0.53 for teams).

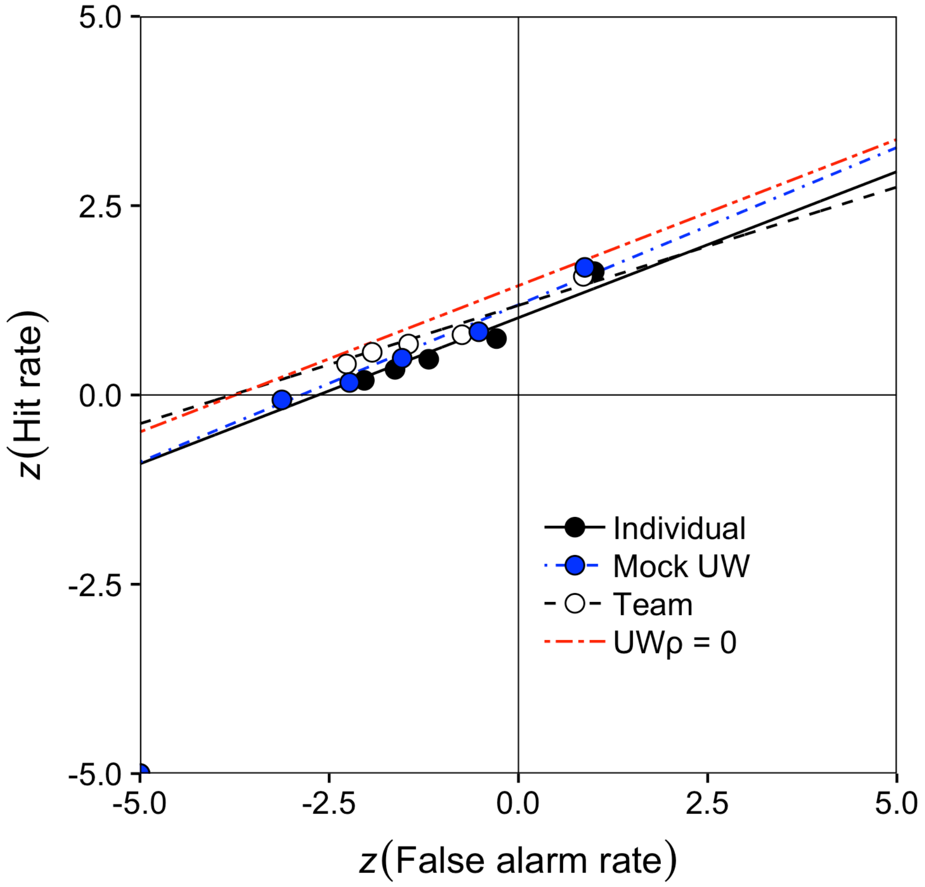

Figure 2.8. *z*ROCs for Experiment 2, with group mean data points superimposed.

Figure 2.9 presents the estimated posterior distributions of d'_e scores, and Figure 2.10 presents distributions of difference scores between observed and predicted team d'_e. Teams (*M* = 1.80, BCI[1.63, 1.98]) again outperformed single observers (*M* = 1.47, BCI[1.35, 1.59]), mean difference = 0.33, BCI[0.12, 0.55]. Team performance fell within the 95% credible interval of the mock UW team scores (*M* = 1.68, BCI[1.51, 1.87]) and fell credibly short of the $UW_{\rho=0}$ model predictions (*M* = 2.08, BCI[1.91, 2.25]).

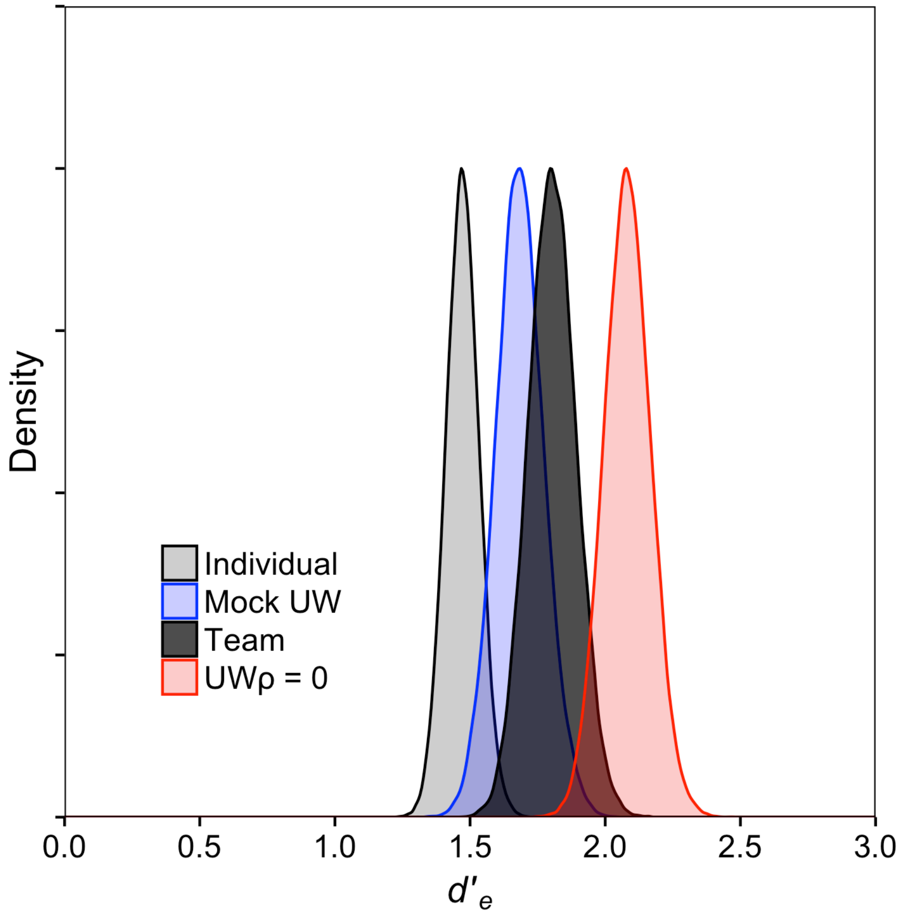

Figure 2.9. Posterior distributions of *d'*$_e$ for single observers (light gray), teams (dark gray), the UW$_{\rho=0}$ model (blue), and mock UW teams (red) in Experiment 2.

Figure 2.11 presents observed and predicted *d'*$_e$ values on a team-by-team basis. The UW$_{\rho=0}$ model overestimated sensitivity for one team (team 4) and underestimated it for another (team 11). Mock UW team sensitivity did not differ credibly from observed sensitivity for any team.

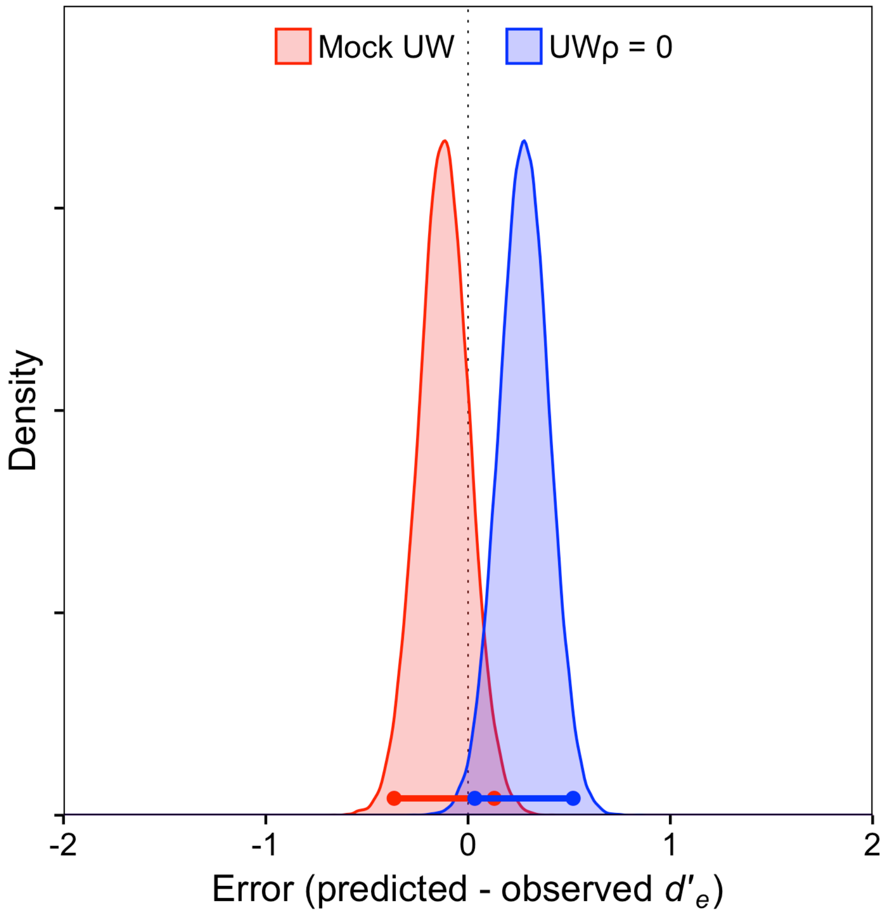

Figure 2.10. Distribution of difference scores between observed and predicted team performance in Experiment 2. Solid lines near the bottom of the figure indicate 95% BCIs.

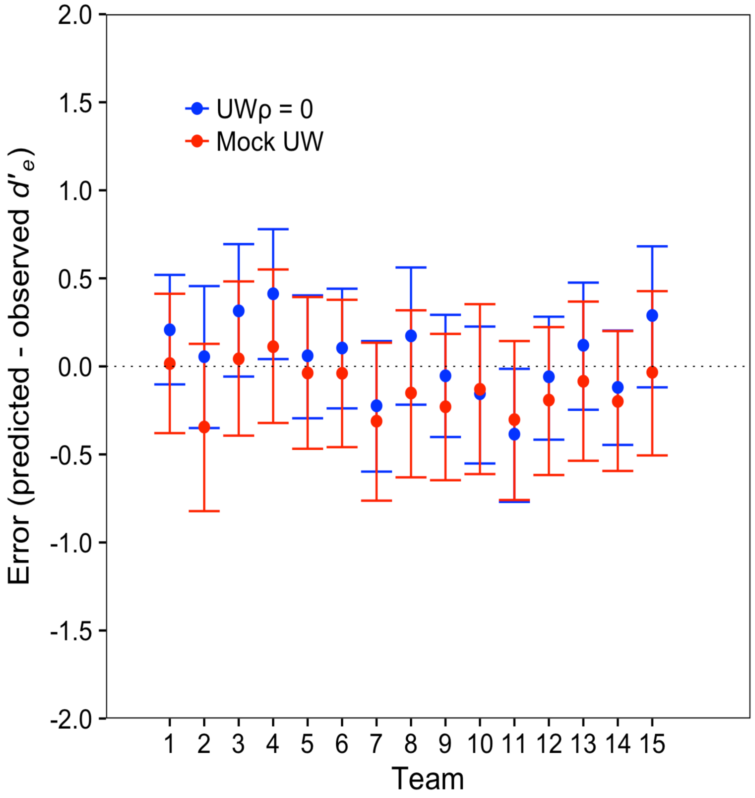

Figure 2.11. Difference scores between observed team *d'e* and $UW_{\rho=0}$ model (blue) and mock UW team (red) predictions in Experiment 2. Error bars are 95% credible intervals on the difference between observed and predicted scores.

Discussion

Team sensitivity fell below the levels predicted by the $UW_{\rho=0}$ model but near the upper end of the distribution of scores for the mock UW teams. These findings are similar to those of Experiment 1 and suggest that any differences in observer motivation between single and collaborative conditions were not directly related to whether single observers and teams completed the baggage search task in the same testing room.

Experiment 3

In some applied contexts, teams may need to jointly inspect a common image from remote locations, communicating electronically. Experiment 3 investigated the effects of collaboration when teams performed both the individual and collaborative conditions in separate testing rooms.

Participants

Sixteen pairs of undergraduate students (23 females, M_{age} = 23.2 *SD* = 6.5) were recruited via the School of Psychology participant pool at Flinders University. Each participant received $20AU in exchange for participation. All participants demonstrated normal color vision and normal or corrected-to-normal visual acuity.

All stimuli and procedures were exactly the same as Experiment 1 except that participants completed both the individual and team conditions is separate testing rooms. In the collaborative condition, team members communicated via Skype (www.skype.com) using only the phone option (i.e., no video). Skype was run on the same testing computers as the visual search task. No Skype window was visible during the search task.

Results

Because of a technical error, data from one observer in the single observer condition were partially lost. All data from that team were therefore excluded from analysis. All of the remaining fifteen teams met the inclusion criterion. Figure 2.12 presents the post burn-in MCMC chains for the model-fitting procedure. By inspection, chains appear to have converged.

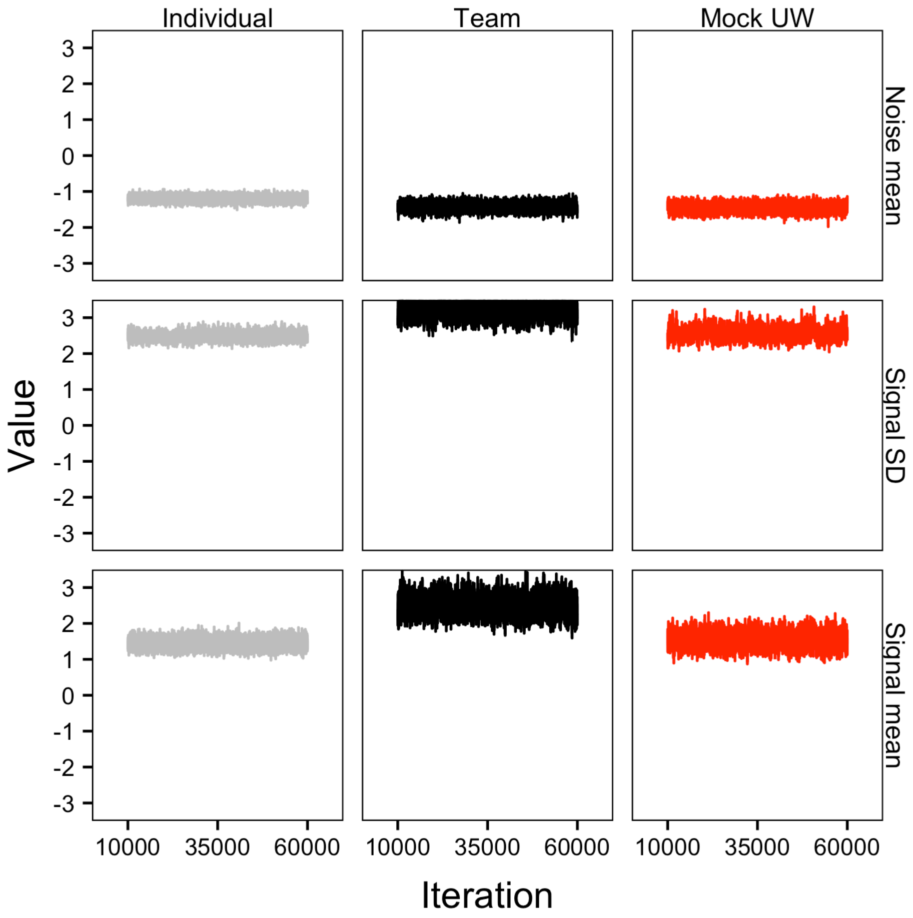

Figure 2.12. MCMC chains for Experiment 3. Columns represent task condition, rows represent estimated parameters.

Figure 2.13 shows the *z*ROCs for the single observers, teams, the UW$_{\rho=0}$ model and mock UW teams, again based on estimates of the population-level parameters. The *z*-slopes for the signal and noise distributions were similar to those of Experiments 1 and 2 (M = 0.54 for single observers, and M = 0.45 for teams) and indicate that the signal distribution had a larger variance than the noise distribution.

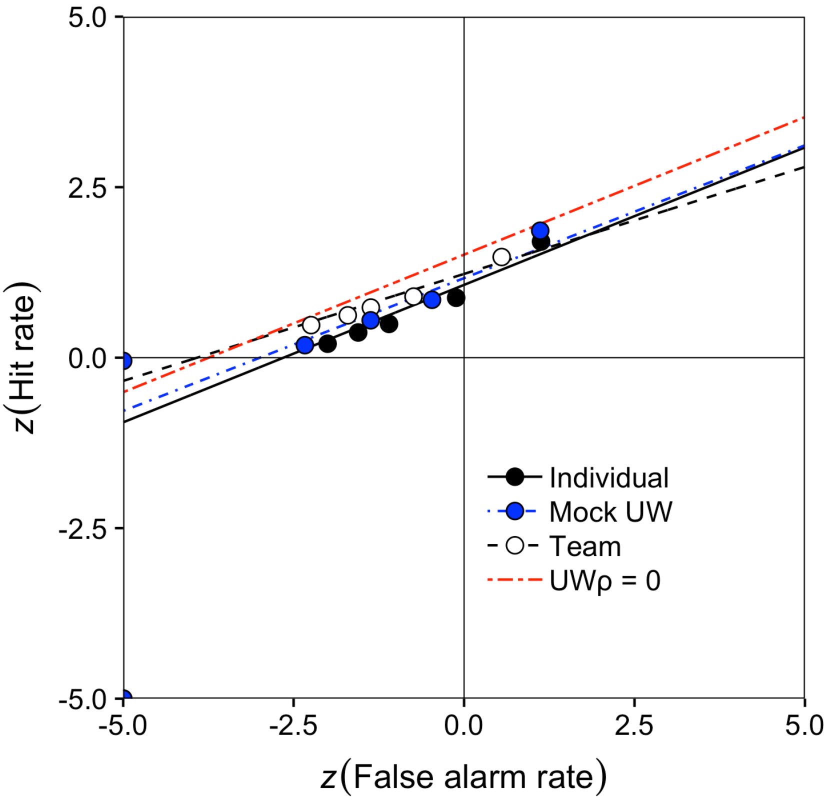

Figure 2.13. *z*ROCs for Experiment 3, with group mean points superimposed.

Figure 2.14 shows the estimated posterior distributions of d'_e scores. Team performance (*M* = 1.87, BCI[1.69, 2.06]) was better than that for single observers (*M* = 1.52, BCI[1.38, 1.66]), mean difference = 0.35, BCI[0.12, 0.58], and fell between that of the mock UW teams (*M* = 1.68, BCI[1.48, 1.88]) and the $UW_{\rho=0}$ model (*M* = 2.15, BCI[1.96, 2.35]).

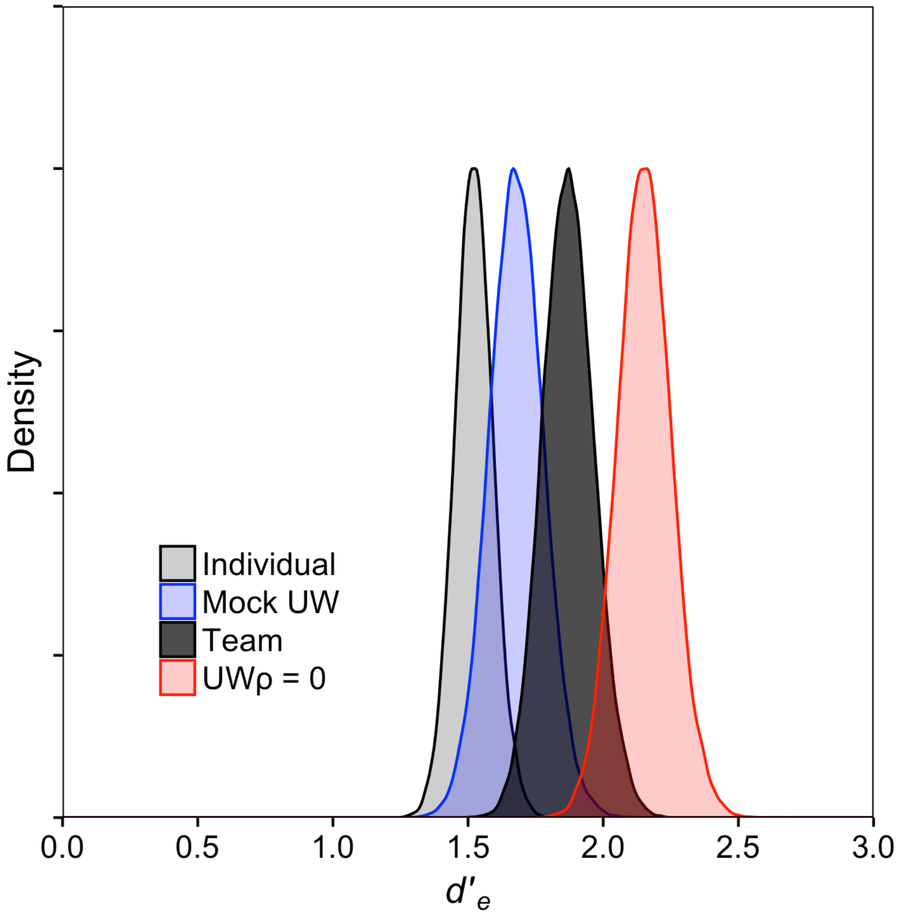

Figure 2.14. Posterior distributions of *d'*$_e$ for single observers (light gray), teams (dark gray), the UW$_{\rho=0}$ model (blue), and mock UW teams (red) in Experiment 3.

The distribution of difference scores between observed and predicted team performances is depicted in Figure 2.15 and shows that observed team performance fell in between that of the mock UW teams and the UW$_{\rho=0}$ model but did not differ credibly from either. A team-by-team comparison of observed and predicted *d'*$_e$ values is presented in Figure 2.16. The mock UW and UW$_{\rho=0}$ model both underestimated sensitivity for one team (team 5).

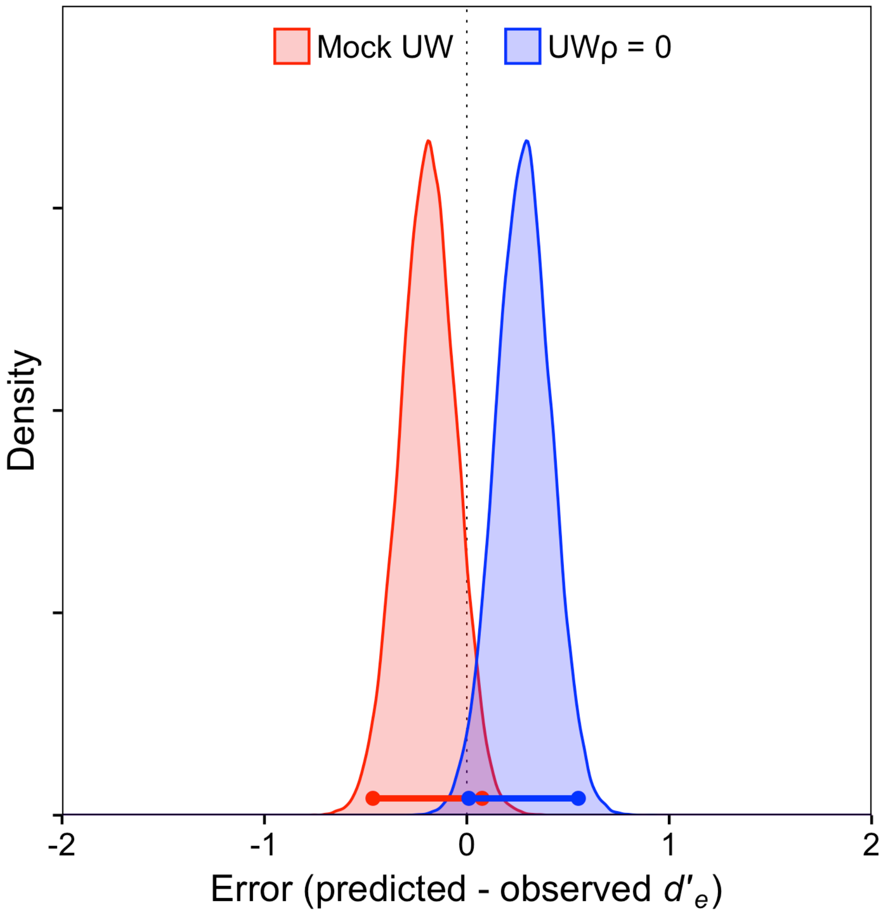

Figure 2.15. Distribution of difference scores between observed and predicted team performances in Experiment 3. Solid lines near the bottom of the figure indicate 95% BCIs.

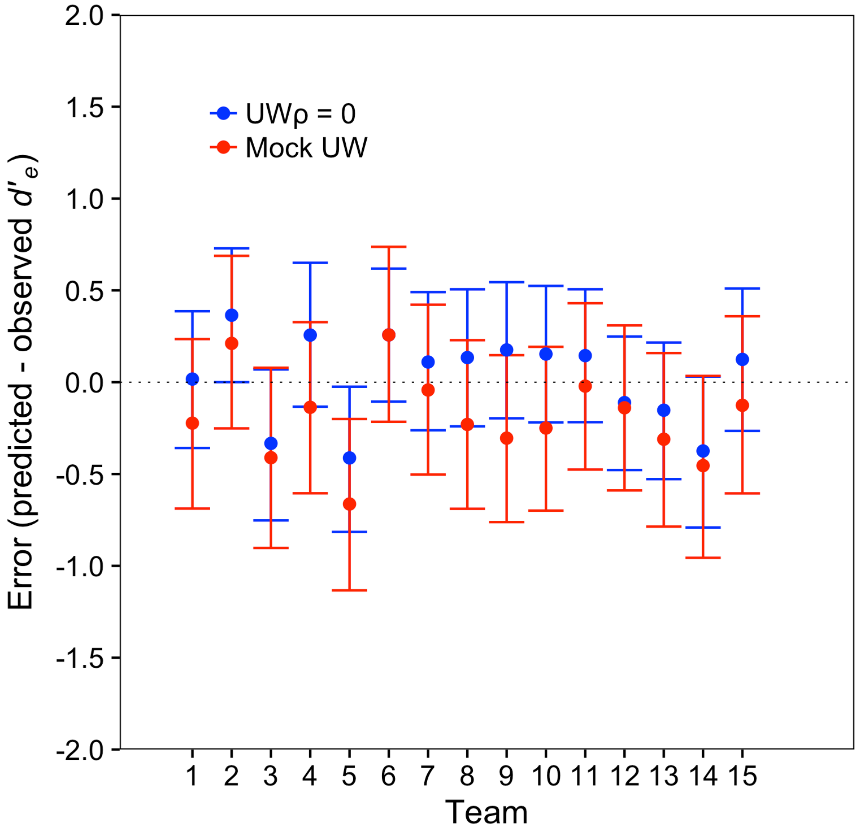

Figure 2.16. Difference scores between observed team *d'ₑ* and $UW_{\rho=0}$ model (blue) and mock UW team (red) predictions in Experiment 3. Error bars are 95% credible intervals on the difference between observed and predicted scores.

Discussion

Team sensitivity exceeded that of single observers and fell roughly midway between performance of the mock UW teams and the $UW_{\rho=0}$ model. This pattern is similar to that seen in Experiments 1 and 2 and indicates that teams need not be collocated to outperform single observers.

Experiment 4

Inspection of response time data from the first three experiments indicated that on average, teams (*M* = 10.02, *SD* = 6.19, *SEM* = 0.39, averaged across experiments) took longer to respond than did single observers (*M* = 3.94, *SD* = 2.06, *SEM* = 0.09). Longer response times in the collaborative conditions presumably reflect, in part, the time needed for team members to discuss their individual judgments and come to a consensus before making a joint response. They may also indicate that the participants took longer to scan images before executing a joint decision, producing a speed-accuracy tradeoff in the search component of the task (Reed, 1973; Wickelgren, 1977) and artifactually inflating team sensitivity relative to the predictions of the two uniform weighting models. Experiment 4 tested this possibility by holding stimulus presentation time fixed at 3s for both the individual and team search conditions.

Participants

Sixteen pairs of undergraduate students (24 females, M_{age} = 23.7 *SD* = 7.8) were recruited via the School of Psychology participant pool at Flinders University. Each participant received $20AU in exchange for participation. All participants demonstrated normal color vision and normal or corrected-to-normal visual acuity.

All stimuli and procedures were identical to Experiment 2 (i.e., participants completed both the individual and collaborative conditions in the same testing room) except that stimulus presentation time each trial was limited to 3s. Stimulus exposure duration was fixed at 3s because this time was slightly less than the mean response time for individual searchers across the first three experiment.

Results

One team was excluded because one member failed to meet the inclusion criterion (d'_e = 0.00), and another was excluded because data for one team member were lost to an evident technical error. This left data from 14 teams for analysis. Figure 2.17 shows the post burn-in MCMC chains for the model-fitting procedure. By inspection, chains appear to have converged.

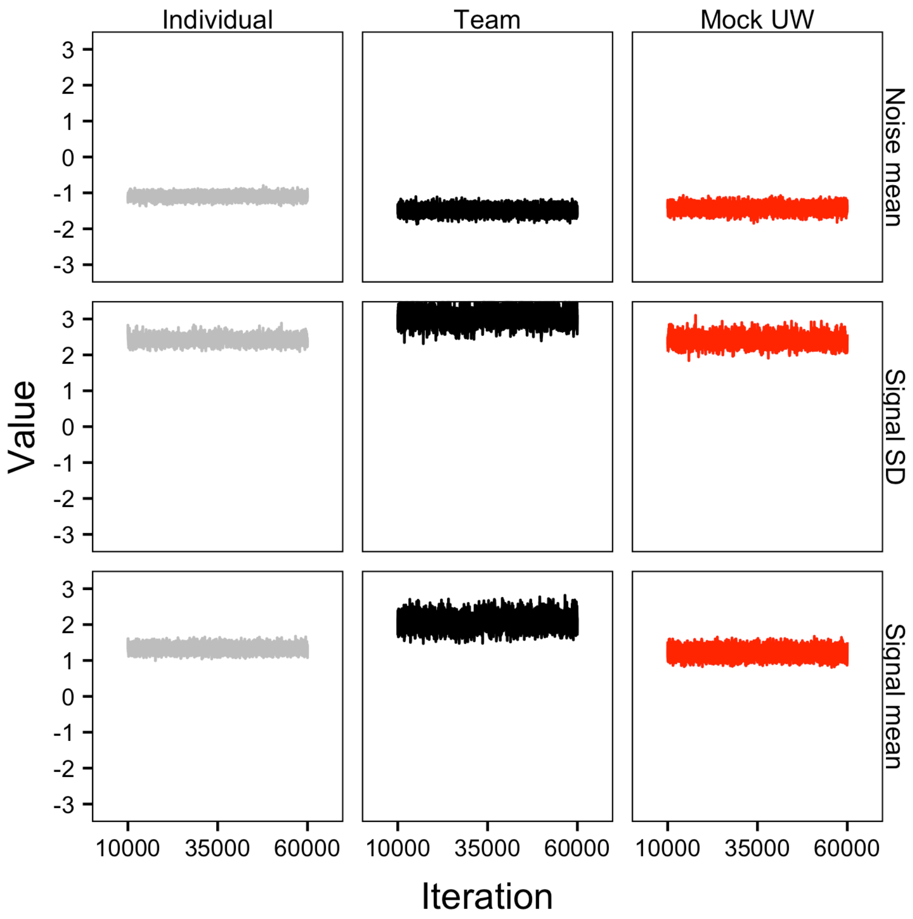

Figure 2.17. MCMC chains for Experiment 4. Columns represent task condition, rows represent estimated parameters.

Figure 2.18 shows the *z*ROCs for the single observers, teams, UW$_{\rho=0}$ model and mock UW team predictions, again based on estimates of the population-level parameters. The *z*-slopes for the signal and noise distributions were similar to those of Experiment 1 ($M = 0.53$ for single observers, and $M = 0.41$ for teams).

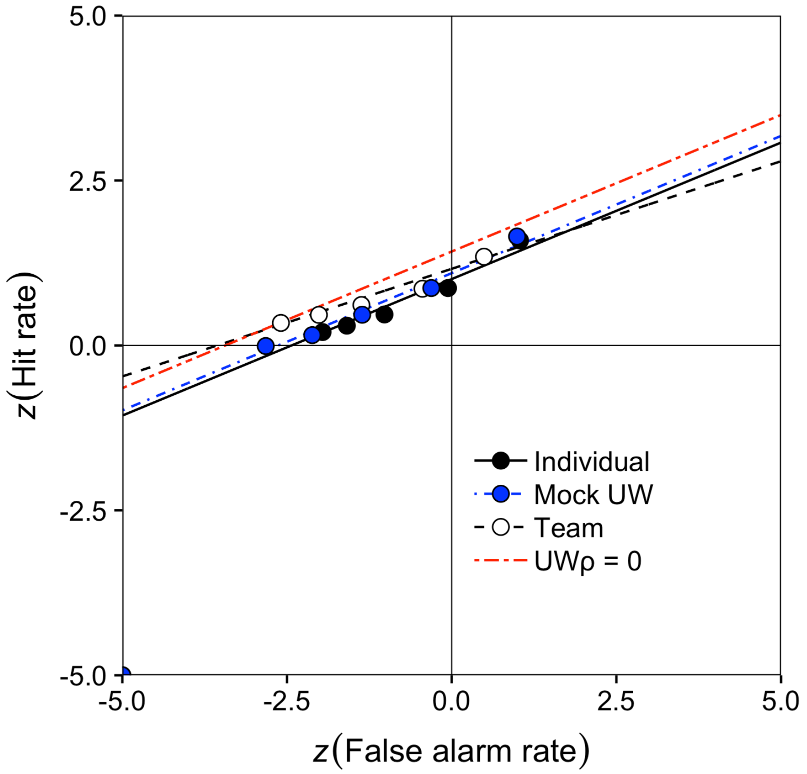

*Figure 2.18. z*ROCs for Experiment 4, with group mean points superimposed.

The estimated posterior distributions of d'_e scores for single observers, team, and the two UW models are presented in Figure 2.19. Distributions of difference scores between observed and predicted team sensitivity are shown in Figure 2.20. Teams (*M* = 1.75, BCI[1.60, 1.91]) outperformed single observers (*M* = 1.42, BCI[1.32, 1.53]), mean difference = 0.33, BCI[0.14, 0.52]. Team sensitivity again fell between that of the UW$_{\rho=0}$ model (*M* = 2.01, BCI[1.86, 2.17]) and mock UW teams (*M* = 1.55, BCI[1.39, 1.71]), though it was not credibly different from mock team performance.

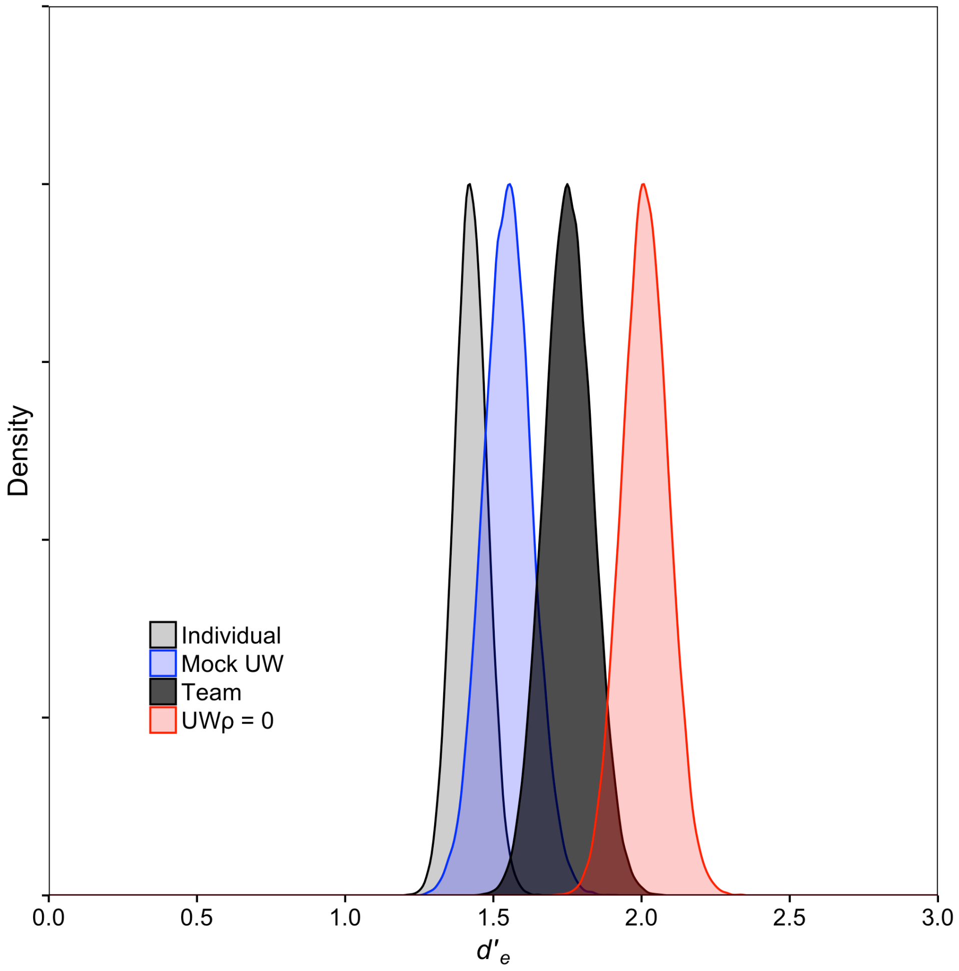

Figure 2.19. Posterior distributions of *d'e* for single observers (light gray), teams (dark gray), the UW$_{\rho=0}$ model (blue), and mock UW teams (red) in Experiment 4.

Figure 2.21 depicts a team-by-team comparison of observed and predicted *d'e*. The mock UW model underestimated sensitivity credibly for three teams (teams 3, 4, and 9). The UW$_{\rho=0}$ model predictions and observed performance not differ credibly for any team.

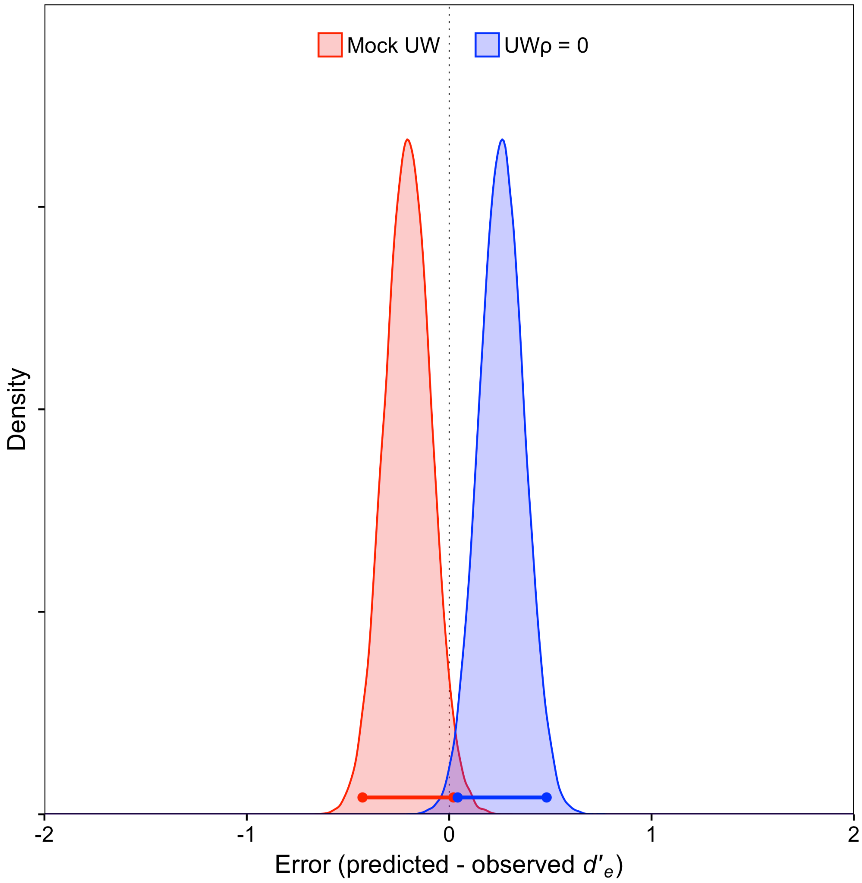

Figure 2.20. Difference scores between observed team performance and $\mathrm{UW}_{\rho=0}$ model (blue) and mock UW team (red) performance in Experiment 4. Solid lines near the bottom indicate 95% BCIs.

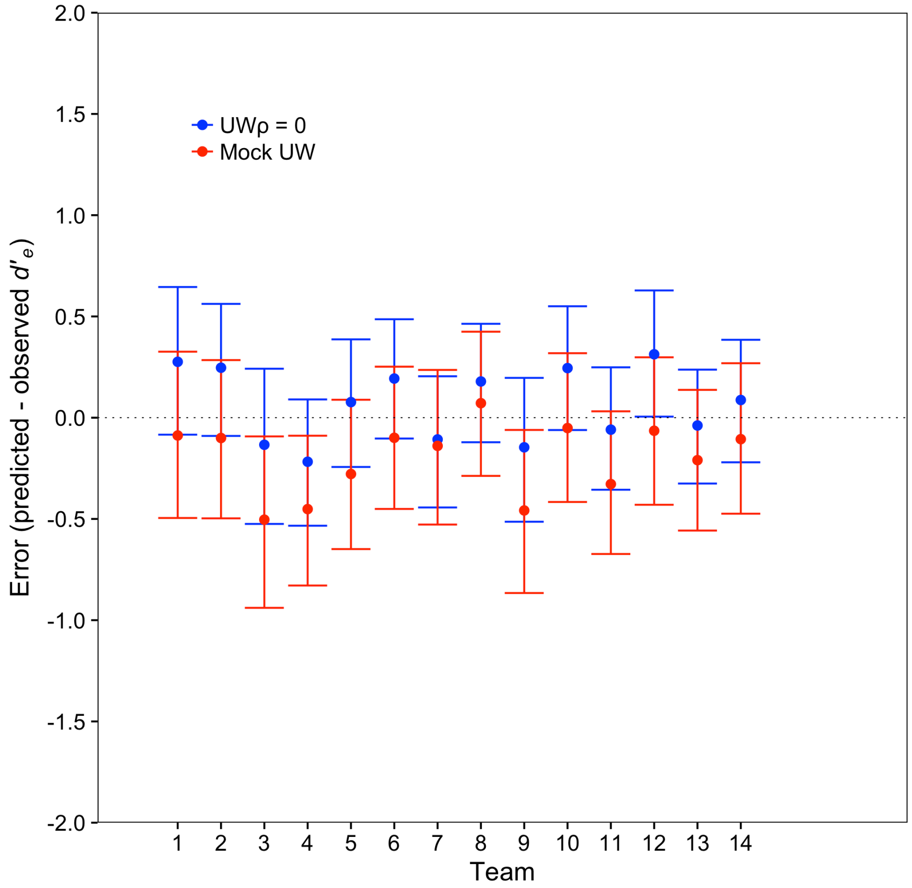

Figure 2.21. Difference scores between observed team d'_e and UW$_{\rho=0}$ model (blue) and mock UW team (red) predictions in Experiment 4. Error bars are 95% credible intervals on the difference between observed and predicted scores.

Discussion

Data showed a pattern of effects very much like that of the first three experiments, despite that viewing time was restricted to three seconds in both the individual and team search conditions. This makes it unlikely that speed-accuracy tradeoffs in visual inspection produced the unexpectedly large team advantage observed in the earlier three experiments.

Meta-analysis of Experiments 1-4

Data across Experiments 1-4 were largely consistent. In all four cases, observed team sensitivity fell roughly midway between the predicted performance of the mock UW teams and the

$UW_{\rho=0}$ model. There was some modest disagreement, however, in the pattern of statistically credible differences across the experiments. In Experiments 1 and 3, observed team performance failed to differ credibly from the predictions of either UW model at the 95% level; in Experiment 2, observed team performance was credibly poorer than expected from the $UW_{\rho=0}$ model performance, but did not differ from performance of the mock UW teams; in Experiment 4, finally, observed team performance credibly exceeded mock UW performance, but did not differ credibly from the $UW_{\rho=0}$ model predictions.

These small inconsistencies across experiments seem likely to reflect statistical variability, rather than genuine differences in team performance levels. Statistical variability is the mostly likely culprit of the small inconsistencies across experiments because of the relatively small sample size, particularly at the group level. To provide a clearer picture of team performance relative to the UW models' predictions, a meta-analysis aggregated the data of the four experiments. The analysis produced a DIC value for the UV, S^2 free of 128825. The MCMC chains for single observers, teams and mock UW teams are presented in Figure 2.22.

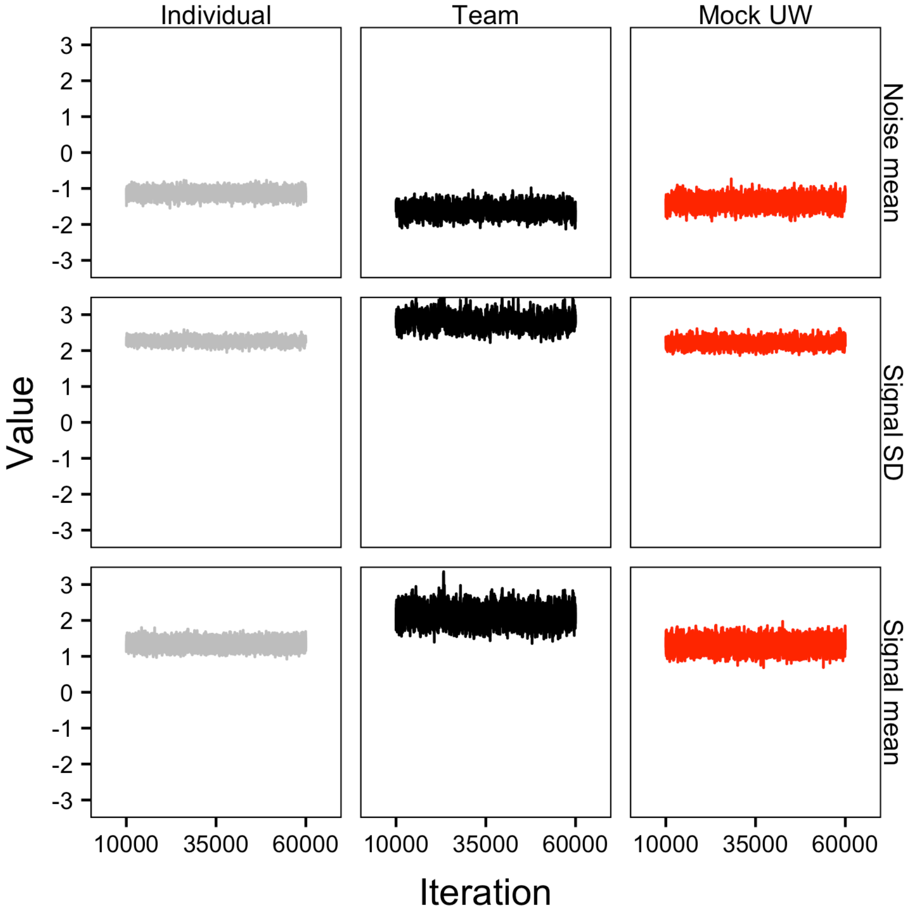

Figure 2.22. MCMC chains for the meta-analysis. Columns represent task condition, rows represent estimated parameters.

Figure 2.23 shows the estimated posterior distribution of d'_e scores, based on estimates of the population-level parameters, for the combined data of all four experiments. Figure 2.24 presents the distribution of difference scores of observed team and UW-predicted team performance. Teams (*M* = 1.85, BCI[1.76, 1.94]) outperformed single observers (*M* = 1.49, BCI[1.42, 1.55]), mean difference = 0.36, BCI[0.25, 0.47], and fell in between that of the mock UW teams (*M* = 1.65, BCI[1.56, 1.74]), and the $UW_{\rho=0}$ model (*M* = 2.10, BCI[2.01, 2.19]).

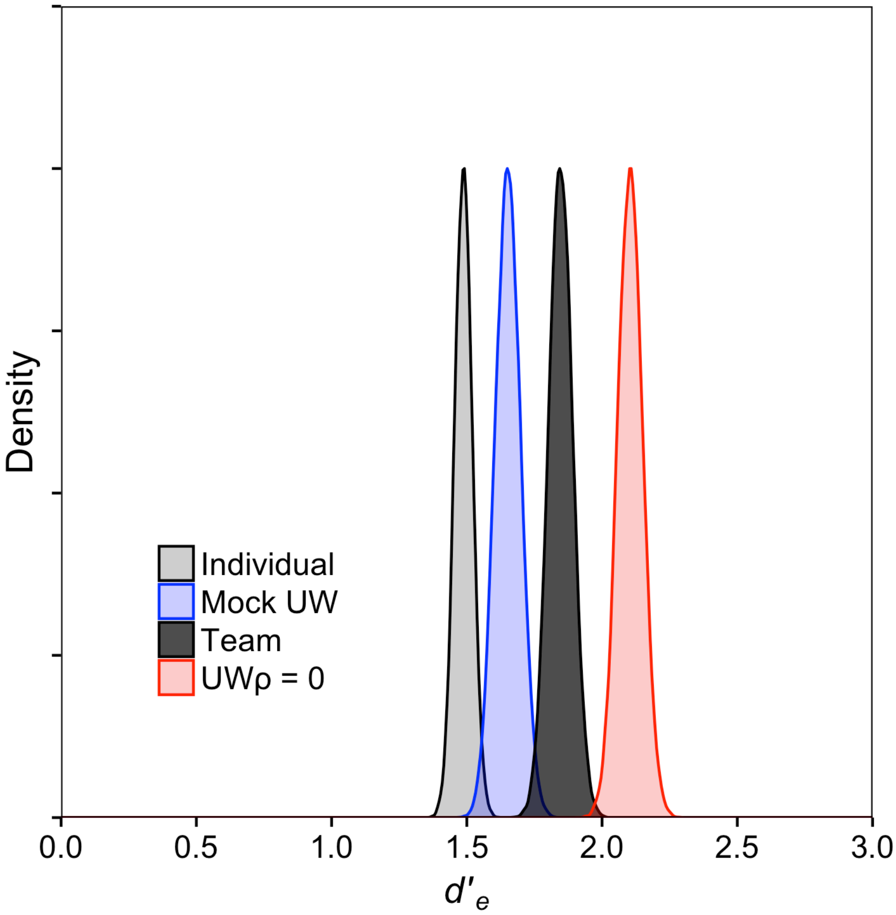

Figure 2.23. Posterior distributions of d'_e for single observers (light gray), teams (dark gray), the $UW_{\rho=0}$ model (blue), and mock UW teams (red) in the meta-analysis.

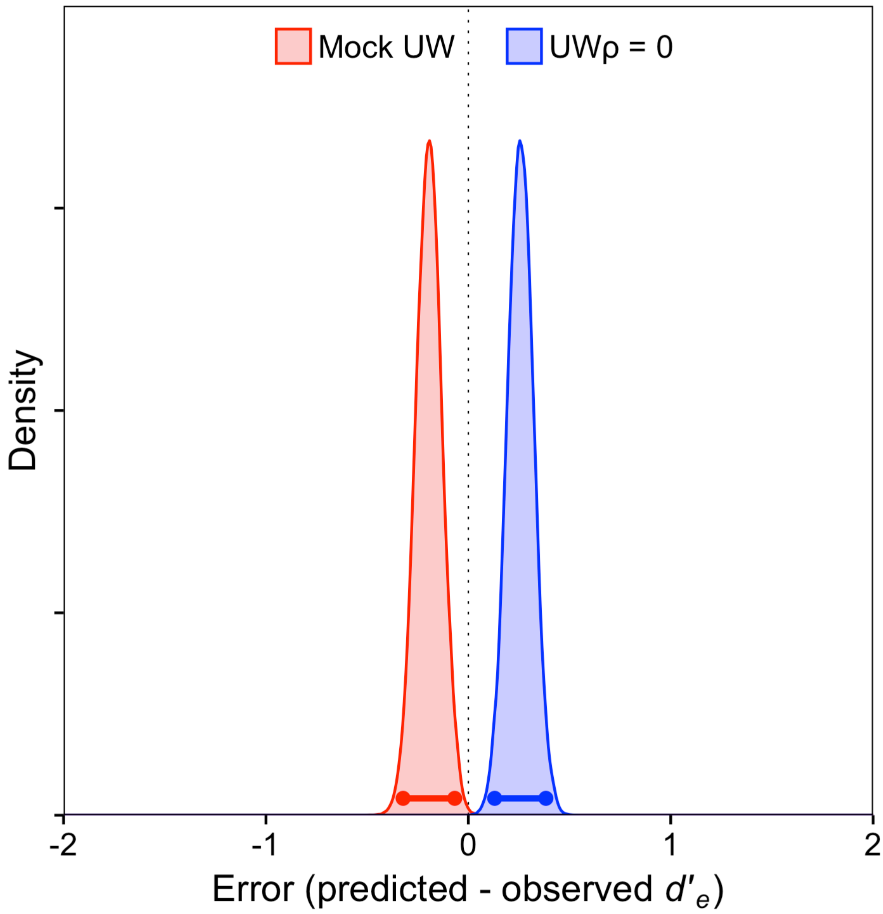

Figure 2.24. Difference scores between observed team performance and $UW_{\rho=0}$ model (blue) and mock UW team (red) performance in the meta-analysis. Solid lines near the bottom indicate 95% BCIs.

Control UW Model

The UW model assumes that participants working as a team communicate and average their raw decision variables in order to reach a decision each trial. In other words, information integration in the standard UW model occurs *before* the observers' raw judgments are converted to discrete choices. In the analyses reported above, however, mock team judgments were calculated by averaging individual team members' confidence ratings on a trial-by-trial basis; information integration occurred *after* the raw judgments had been discretized. This raises the concern that by discarding information available in the raw judgments, analysis of the mock team judgments may have underestimated the performance expected from a UW strategy.

A control analysis tested this possibility by adopting a converging method of estimating UW performance for stochastically dependent team members (Metz & Shen, 1992). In this approach, we take the correlation between team members' confidence ratings as an estimate of the correlation between their unobservable decision variables, use that value to estimate the covariance of the team members' decision variables, then adjust the variance of the predicted noise and signal distributions to incorporate that covariance. If r_n and r_s are the correlations between participants confidence ratings on noise and signal trials, respectively, then the covariance between the team members' raw decision variables

$$\mathrm{cov}\left(X_{n_1}, X_{n_2}\right) = r_n$$

for noise trials, and

$$\mathrm{cov}\left(X_{n_1}, X_{n_2}\right) = r_s \sigma_1 \sigma_2$$

for signal trials. The noise distribution for the team decision variable becomes,

$$X_{n_1+n_2} \sim N\left(0, 2 + 2 \times \mathrm{cov}\left[X_{n_1}, X_{n_2}\right]\right),$$

and the signal distribution for the team decision variable becomes,

$$X_{s_1+s_2} \sim N\left(\mu_1 + \ \mu_2, \sigma_1^2 + \sigma_2^2 + 2 \times \mathrm{cov}\left[X_{s_1}, X_{s_2}\right]\right).$$

These distributions specify the ROC for a UW team whose members judgments are correlated, from which *d'e* can be calculated. We will refer to this as the $UW_{\rho > 0}$ model.

To calculate predictions for this model, we first aggregated confidence ratings across the full set of 60 teams included in the meta-analysis above, arbitrarily identifying one participant within each team as observer 1 and the other as observer 2. We then used the model described by Kruschke (2017) to produce a Bayesian estimation of the correlation between team members' confidence ratings. Estimates were based on a total of 50,000 MCMC iterations (four chains of 12,500 iterations each) following a burn-in sequence of 10,000 iterations. This produced estimates of $r_n = 0.14$ and $r_s = 0.46$.

These values imply that dependencies in team members' judgments were driven largely by visibility of the target, rather than by characteristics of the image clutter.

Figure 2.25 presents the estimated posterior distribution of team *d'e* scores based on the $UW_{\rho > 0}$ model, along with the estimated posterior of mock UW team scores. The two methods produced similar distributions, lending confidence that the mock UW teams scores did not dramatically underestimate the sensitivity levels expected from a UW strategy. Note that because r_n and r_s are based on discrete confidence ratings, they are likely to underestimate the true correlation between the team members' raw judgments and might therefore cause the $UW_{\rho > 0}$ model to overestimate sensitivity. This again lends confidence that the differences between mock UW team sensitivity and observed team sensitivity were not artefactual.

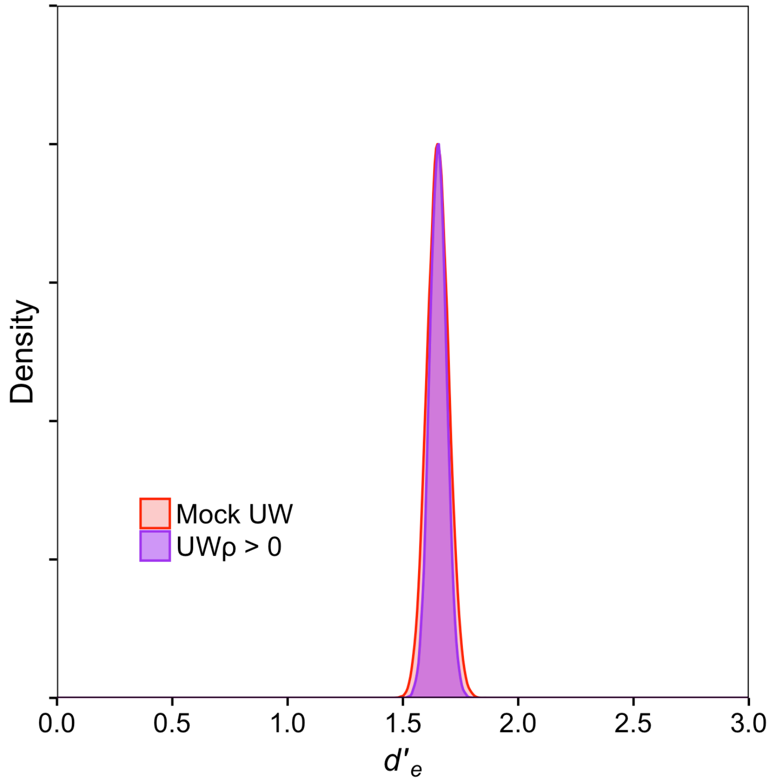

Figure 2.25. Posterior distribution of *d'*$_e$ for mock UW teams (red) and the $UW_{\rho > 0}$ model (purple) performance for Experiments 1-4.

General Discussion

In Experiment 1, we asked single observers to complete the task in separate testing rooms, whereas in Experiment 2, single observers completed the task in the same testing room. Experiment 3 simulated a non-co-located joint search in which both single observers and teams completed the search task in separate testing rooms. Finally, single observers and teams in Experiment 4 were located in the same testing room and stimulus presentation time was limited to 3s. In all four experiments, teams performed better than single observers, and achieved sensitivity levels roughly midway between the performance of mock UW teams and a UW model assuming stochastically independent observers. A meta-analysis using the data from all four experiments confirmed this pattern of effects.

These results echo those of Malcolmson et al. (2007), who also found that empirical teams outperformed mock teams based on a UW strategy, but differ from those of Bahrami et al. (2010), Bahrami et al. (2012) who report collaborative performances that approach levels predicted by the UW model in simple perceptual and cognitive tasks. Their experiments employed tightly controlled stimuli to ensure uncorrelated judgments from team members as well as equal-variance noise-alone and signal-plus-noise distributions. The current data show performance slightly above the UW level in a more naturalistic task, in which judgments between observers were not stochastically independent and the assumption of equal-variance evidence distributions was violated.

The finding that teams achieved performance levels better than expected given their correlated responses implies that participants overcame the negative effect of correlated team member responses, though the data do not reveal how they achieved this. One possibility is that individuals may have increased their effort, and thus sensitivity levels, during collaboration to avoid negative performance evaluation. Although the phenomenon of social loafing is perhaps more familiar, the opposite effect, social facilitation, is also possible (Kerr & Tinsdale, 2004). Social comparison, for example, can motivate team members to work harder under groups conditions than performing a task individually (Stroebe, Diehl, & Abakoumkin, 1996; Weber & Hertel, 2007). This implies that teams could have outperformed mock teams because either or both members of a team put forth more effort under collaborative conditions than they did working individually. This suggestion may seem to sit poorly with the finding in Experiment 2 that individual observers did not seem to put forth additional effort when working in the same testing room compared to when they worked in different locations (Experiment 1). When single observers worked in the same testing room, however, they were instructed to keep their eyes to their own computer displays, to refrain from communicating in any way, and to exit the room upon completing the task. As such, they would have had little chance to compare their own performance to their partner's and might not have felt the same social pressure to excel in the individual search conditions that they did in the team conditions.

Collaborators could also have boosted their performance by exchanging information in a way that allowed them to sample the stimulus images more effectively or extensively. The OW and UW models discussed above, notably, assume that collaborators integrate their individual judgments to reach a team decision, but that collaboration does change the process by which the team members sample information from the stimulus. Earlier work has often enforced this constraint by generating stimuli independently for each team member and presenting them in isolation (e.g., Bahrami et al., 2010; Sorkin et al., 2001). Presenting a common stimulus for inspection, however, allows collaborating team members to guide or inform each other's sampling strategies, as might happen, for instance, if one collaborator points the other toward a suspicious item within the stimulus that the second might otherwise not have inspected. In effect, collaborating over a common stimulus would transform a collaborative team from a standard parallel system to an interactive parallel system (Mordkoff & Yantis, 1991; Townsend & Wenger, 2004). Interacting channels systems are dramatically more efficient that standard parallel systems (Eidels, Houpt, Altieri, Pei, & Townsend, 2011; Townsend & Wenger, 2004), and thus could extract greater amounts of evidence from a stimulus even within a fixed sampling period.

One form of interactive channels operation that participants might have adopted is a division-of-labor strategy. Earlier studies of collaborative search have found that team members tend to divide responsibilities, for example, by allocating each team member a different region of the display for inspection (Brennan et al., 2008; Chen, 2007; Malcolmson et al., 2007), or by allocating each team member a different target to search for (Chen, 2007). Paired searchers in the current experiments might likewise have focused attention on different regions of the stimulus images or adopted attentional sets for different targets. These strategies could have been adopted either purposefully or tacitly (Chen, 2007).

The benefits of a division-of-labor strategy would reflect a tradeoff among three effects. Consider a search task with four potential target items, A, B, C, and D. By restricting their attention to a subset of the search space—defined by literal space within the display and the space of potential target items—searchers would sacrifice sensitivity for items outside the attended space. A searcher with an exclusive attentional set for Target A would lose sensitivity for detecting targets B, C, and D (Chen & Zelinsky, 2006; Schmidt & Zelinsky, 2009, Vickery, King, & Jiang, 2005), for instance, and a searcher whose attention was focused on one region the display would be less likely to notice a target embedded in clutter elsewhere (Kundel, Nodine, & Carmody, 1978; McCarley, Kramer, Wickens, Vidoni, & Boot, 2004). However, restricting attention to a narrow subset of the search space would improve sensitivity for items within the attended subspace. Search performance is more efficient when observers search for one target at a time, instead of multiple simultaneously (Menneer, Barrett, Phillips, Donnelly, & Cave, 2007; Menneer, Cave, & Donnelly, 2009), for example.

Finally, by attending to different regions or characteristics of a stimulus image, collaborators using a division-of-labor strategy could potentially de-correlate their judgments. A searcher attending to the left half of the display might make a low-confidence target-absent judgment, for instance, while her partner attending to the right half of the display detects the target with high confidence. A reduction in the correlation between judgments, as discussed above, would tend to increase team sensitivity.

Future research, perhaps using eye-tracking or the analysis of collaborators' utterances, might reveal which of these strategies, if any, collaborators in the current task used to bolster performance above levels of the mock UW teams.

Constraints on generality

A number of constraints on generality (Simons, Shoda, & Lindsay, 2017) of the current findings should be noted. Participants completed the task in a quiet environment with no distractions, unlike many real-world baggage screening environments. Also, participants comprised non-expert screeners. Future research is needed to confirm if the obtained pattern of results generalise to expert screeners.

It is also worth noting that many real-world situations, including airport security, are likely to entail an exceptionally low (≤ 1%) rate of target prevalence, i.e., frequency with which targets are presented (Wolfe et al., 2007). Target prevalence in all four of the current experiments was 40%, and so a lower target prevalence would have better mimicked naturalism but also increased the potential for a low-prevalence effect, when targets are more often missed due to their low probability (Wolfe et al., 2007). In an x-ray baggage signal detection task, Wolfe et al. (2007) examined the sensitivity and miss errors of paired observers across low (2%) and balanced (50%) target prevalence conditions and expected the miss rate of pairs to be the product of the two individual team member's miss rates. Unexpectedly, paired observers demonstrated a low-prevalence effect worse than expected from the product of their individual miss rates and only slightly lower than the better of the two observers. Importantly, poor performance in the low target prevalence condition was characterized by a shift in criterion and not a decrease in sensitivity. In fact, teams achieved greater sensitivity levels in the low prevalence condition than in the equal prevalence condition, but maintained a strong bias toward responding "no," resulting in very low hit rates. In other words, collaboration might improve sensitivity, but without buffering against performance lapses caused by suboptimal criterion-setting.

This concludes the currently available paper.

CHAPTER 3:

STUDY 2

The following manuscript entitled, *Collaborative searchers outperform individuals in the absence of precise target information*, is presently available on PsyArxiv Preprints (https://doi.org/10.17605/OSF.IO/V6CQK). The version of the manuscript presented here is the same version currently available.

Both authors were involved in the formulation of the study concept and design, and data analysis. Ali Enright collected the data and completed the initial draft of the manuscript. Jason McCarley edited multiple revisions of the manuscript.

Abstract

Two-person teams outperform individuals in a simulated baggage x-ray screening task, and even appear to exceed expectations based on statistical limitations (Enright & McCarley, 2018). The current experiments aimed to replicate and extend this result. We use Bayesian hierarchical modelling of receiver operating characteristics to examine collaborative visual search performance in a visual search task wherein top-down target information was constrained. Participants (*N* = 32, 16 teams per experiment), working independently or collaboratively, performed a visual search task framed as a medical image reading task. Stimuli were polygons generated by randomly distorting a prototype shape. Each trial, observers judged whether an extreme distortion was present among a set of low-distortion distractor objects. Team members' individual sensitivity levels were used to predict collaborative sensitivity using two versions of a uniform judgment weighting model, one assuming stochastically independent judgements from the two team members and the other accounting for correlations in the team members' judgements. Collaborative search performance was better than that from single observers in both Experiment 1 and 2 and fell roughly midway between the predictions of the correlated and uncorrelated models. Results imply that collaboration when searching in conditions of limited top-down target knowledge is beneficial.

Introduction

Signal detection theory (SDT; Green & Swets, 1966; Macmillan & Creelman, 2005) provides a framework for studying decision makers' ability to reach discrete judgments from uncertain data. The prototypical signal detection task asks observers to distinguish two states of the world, one termed signal-plus-noise and the other noise-alone, from probabilistic evidence. In the standard SDT model, the signal-plus-noise and noise-alone evidence distributions are normal with different means but the same standard deviation (Macmillan & Creelman, 2005). *Sensitivity* denotes the ability to discriminate signals from noise, and is measured by *d'*, the distance between the means of the signal-plus-noise and noise-alone distributions, in standard deviation units (Green & Swets, 1966; Macmillan & Creelman, 2005).

SDT also offers models for understanding collaborative or team decision making. Collaborative sensitivity in a signal detection task reflects individual team members' sensitivity levels, their information integration strategy, and the correlation between their judgments (Bahrami et al., 2010, 2012; Sorkin & Dai, 1994; Sorkin, Hays, & West, 2001; Sorkin, West, & Robinson, 1998). The *Optimal Weighting Model* (OW; Bahrami et al., 2010; Sorkin & Dai, 1994; Sorkin et al., 2001) predicts ideal collaborative sensitivity. Within this model, a team reaches a decision by combining team members' individual judgments, weighting them according to each individuals' mean sensitivity. Assuming an equal-variance Gaussian model and stochastic independence between team members' judgments, *d'* for the group is,

$$d'_{OW} = \sqrt{\sum d'^2_i}. \qquad [1]$$

The *Uniform Weighting Model* (UW; Sorkin & Dai, 1994) is similar, but assumes that team members' judgments are weighted equally when averaged to reach a team decision. Group *d'* under this model is,

$$d'_{UW} = \frac{\sum d'_i}{\sqrt{N}}. \qquad [2]$$

When team members are equally sensitive, the UW model is equivalent to the OW model. When team members are not equally sensitive, the UW model predicts lower sensitivity than the OW model.

Studies of collaborative signal detection have consistently found that groups outperform individuals, but at varying levels of efficiency. Hinsz (1990) demonstrated that 6-member-groups achieved higher sensitivity than individuals when recalling audiovisual information but performed far below the predictions of the UW/OW model. In contrast, Bahrami and colleagues (Bahrami, Olsen, Bang, Roepstorff, Rees, & Frith, 2012; Bahrami, Olsen, Latham, Roepstorff, Rees, & Frith, 2010) examined team performance in a 2-interval forced-choice visual search for a contrast singleton and found sensitivity levels similar to the predictions of a UW model. This pattern obtained even when collaborators differed dramatically in their individual sensitivity levels, meaning that the UW strategy led to performance worse than the better team member could have achieved alone. Sorkin et al. (2001) found even better performance in a multiple-cue judgment task, showing that small groups (4 or fewer members) performed with near-perfect efficiency, approaching predictions of the OW model. Efficiency decreased as group size increased, but this appeared to result from social loafing, rather than inefficient weighting strategies.

In the experiments conducted by Bahrami et al. (2010, 2012) and Sorkin et al. (2001), importantly, stimuli were generated independently for each team member each trial, and team members did not inspect each other's stimuli. These controls minimized the correlations between participants' judgments, consistent with the assumption of stochastically independent collaborators on which Equations 1 and 2 rest. Team sensitivity is reduced when team members provide correlated judgments (Sorkin et al., 2001), as is likely to be the case when collaborators make judgments of a common stimulus. Consider the example of two pathologists jointly searching a cell sample for an abnormal cell (Dee, 2009; Dee et al., 2003). Although unique variance in their judgments might result from differences in the observers' sensory abilities, oculomotor scan patterns, background knowledge, etc., ambiguities in the image itself will provide a strong source of stochastic dependency between the observers (Sorkin & Dai, 1994; Sorkin et al., 2001). The predictions of the OW and UW models can be adjusted to account for stochastic dependencies between observers (Sorkin et al., 2001), but this requires that the correlation between the observers' judgment be known. Unfortunately, because the observers' mental decision variables are unobservable, the correlation between them is difficult to estimate (Metz & Shen, 1992).

An alternative method for generating predictions for correlated observers is to create nominal or mock teams by combining team members' individual judgments for yoked stimuli (Metz & Shen, 1992). The mock team judgments, because they are based on isolated individuals' responses to the same stimuli, inherently account for stimulus- or sequence-driven dependencies in the decision makers' judgments. Mock team judgments, in other words, reflect the collaborative sensitivity that can be expected given the stochastic dependency between the team members' judgments. Malcolmson, Reynolds, and Smilek (2007) compared empirical and mock team sensitivity in a visual search task. Two-person teams completed the task working together (empirical teams) and alone in separate testing rooms (mock teams). In the latter condition, both members of the team experienced the same sequence of trials, and team judgments were produced by combining the individual members' yes-no judgments using a disjunctive rule. Empirical teams produced greater sensitivity levels than did mock teams, suggesting performance better than expected from a UW rule given the statistical dependencies in individual team members' judgments. Participants reported informally that their collaborative strategy was to divide the display for search, with one team member attending one half of the search field and second team member attending to the other half.

More recently, Enright and McCarley (2018) examined collaborative performance in a simulated baggage screening task. Participants viewed a series of baggage x-rays and judged whether a knife was present each trial. In individual search conditions of Experiment 1, the participants performed the task in isolated rooms. In the collaborative conditions, they sat side-by-side and were allowed to discuss the stimulus before making a team judgment. Because the stimuli were expected to violate the assumption of equal-variance signal and noise distributions, participants were asked to provide confidence ratings in place of simple yes-no judgments, and data were used to plot receiver operating characteristics (ROCs; Macmillan & Creelman, 2005). Observed ROCs were compared to the predictions of two versions of the UW model. The first, denoted the *mock UW* model, assumed stochastically dependent judgments between team members, and the other, denoted the $UW_{\rho = 0}$ model, assumed independent judgments. Mock UW team predictions were derived by averaging individual team members' ratings of yoked stimuli to create team judgments. The $\mathrm{UW}_{\rho = 0}$ model predictions were adapted to predict collaborative signal detection performance in ROC space without the assumption of equal-variance distributions. Sensitivity was gauged with the statistic *d'ₑ* (Egan, Schulman, & Greenberg, 1959; Macmillan & Creelman, 2005), a generalization of *d'*, derived from analysis of the ROC, that does not require the assumption of equal-variance evidence distributions.

Echoing the findings of Malcolmson et al. (2007), collaborative sensitivity was better than that from single observers, and fell in between the predictions of the UW and mock UW model predictions. This pattern of effects persisted when participants performed the individual search task while sitting at separate workstations in the same room, when they performed the collaborative search task from different rooms while communicating via speakerphone, and when viewing times were restricted. Two interpretations of the results were considered. The first was that the results reflected social compensation effects suggesting teams put forth more effort when working collaboratively than independently. The second interpretation was that teams' might have interacted while viewing the images in a way that improved their scanning or information sampling. Most obvious was the possibility that searchers adopted a division-of-labor strategy akin to that of Malcolmson et al.'s (2007) participants, making individual team members responsible for inspecting particular regions of the display or detecting particular targets (see Brennan et al., 2008, Chen, 2007, for evidence of similar strategies in speeded collaborative search).

The present experiments aimed to extend the findings of Enright and McCarley (2018), in two important ways. The first goal was simply to replicate those earlier findings using an alternative stimulus set. Although Enright and McCarley (2018) found consistent results across a series of five experiments, the stimuli, simulated baggage x-rays with knives as targets, were the same in all cases. This raises the concern that the observed results might have been idiosyncratic to the stimuli. The second goal of the current experiments was to replicate our earlier findings in a task designed to limit target certainty. In our previous experiments, participants searched for targets drawn from a set of five knives. Pictures of the five potential targets were provided with the task instructions, and the participants performed a set of practice trials that gave them further exposure to the targets before beginning the experimental trials. Thus, although participants did not have perfect certainty of which target might appear on any given trial, the space of potential target items was small and familiar.

However, many naturalistic tasks require observers to judge whether a stimulus contains a target whose appearance is uncertain. As in the pathologist example above, the cell sample might contain an anomalous cell whose shape, size, and color are not predetermined. Some theories of visual search (e.g., Wolfe, 1994) posit that top-down knowledge of the target helps guide visual search to locations on an activation map likely to contain target features. Visual search performance without top down guidance limits attentional guidance and is generally less efficient than search with good top-down guidance (Chen & Zelinsky, 2006; Wolfe, 1994). Search is more efficient if observers are provided a detailed and accurate visual representation of the target, for example, than if they are provided with a text description or an imprecise or degraded visual representation (Hout & Goldinger, 2015; Malcolm & Henderson, 2009; Schmidt & Zelinsky, 2009; Vickery, King, & Jiang, 2005).

Smith and colleagues (Smith, Redford, Gent, & Washburn, 2005) found particularly poor performance in a novel form of visual search characterized by weak top-down control. Stimuli were randomly-generated polygons created by distorting prototype shapes (Posner, Goldsmith, & Welton, 1967). Participants searched each trial for shapes derived from a designated set of prototype objects, amongst distractors that were not derived from the target prototypes. In most cases, targets were high-level distortions of the prototype. Target uncertainty was therefore high and top-down control necessarily poor. Remarkably, sensitivity under these conditions was near chance levels. Performance substantially exceeded chance only when targets were presented without distractors, or when targets were highly similar to their category prototypes, providing high target certainty (Smith et al., 2005).

The current experiments adapted the task and stimuli of Smith et al. (2005) to introduce further target uncertainty, limiting the prospects of a divide-and-conquer collaborative strategy. Here, all the objects within a given stimulus image were distortions of a common prototype. The distractors, though, were low-level distortions, and the target, when it was present, was a high-level distortion. Thus, the target could be distinguished only by comparison to the surrounding distractors. As noted, participants in Enright and McCarley's (2018) experiments could have divided responsibility in either of two ways, either by searching different regions of the display, or by searching for different items within the set of five potential targets. The current stimuli limit both these strategies. Without foreknowledge of the target shapes, collaborators could not adopt an attentional set for any particular target or adopt a strategy of searching for different targets. And because the target was defined as a distortion more extreme than the surrounding items, collaborators could not identify any single item as a target without also attending to the distractors. Rather, they were required to attend to all stimuli presented in the search display. Comparing search stimuli in this way inherently limited the ability of team members to restrict search to predefined search areas.

Building from Enright and McCarley (2018), we use Bayesian hierarchical ROC analysis to examine collaborative visual search using dot-distortion stimuli. The search was framed as a medical image reading task. Participants searched cell samples for an abnormal cell, working collaboratively, in teams of two, or independently. Graded confidence responses were collected to allow analysis of the ROC. Empirical collaborative performance was compared to the predictions of two versions of the UW model, the mock UW model and the $UW_{\rho = 0}$ model.

Experiment 1

In Experiment 1, participants performed the visual search task individually, in separate testing rooms, and collaboratively, sharing one computer in the same room. Experimental methods were preregistered (Enright, McCarley, & Leggett, 2018a, June 13) (https://osf.io/u43yg/?view_only=78d0a609a9b64dcb8d1b9c70094c6dc3).

Method

Participants

Sixteen pairs of undergraduate students (22 female, M_{age} = 22.5, SD = 3.28) were recruited via Finders University College of Education, Psychology, & Social Work's first year research participant pool. All participants demonstrated normal or corrected-to-normal visual acuity (tested using a standard eye chart in the lab) and colour vision (determined using Ishihara test) and were paid $20AU in exchange for participation.

Apparatus and Stimuli

Participants completed the visual search task on a 370 mm x 300 mm Samsung monitor (model S24D590PL), with a resolution of 1920 x 1080 pixels and a refresh rate of 85 Hz. Stimulus display and response collection were controlled by software custom written in PsychoPy (Peirce, 2007, 2009). Participants viewed displays from a distance of roughly 570mm, however viewing distance was not constrained.

Dot-distortion stimuli (see Figure 3.1 for example stimuli) were generated in RStudio (www.rstudio.com). Each stimulus image comprised a set of 3-5 polygons created by distorting a common prototype. A prototype was generated by randomly selecting a sequence of five points within a 30 x 30 (21° x 21°) grid, then connecting them in order. Distortions were created by randomly displacing the prototype's vertices (Posner et al., 1967; Smith et al., 2005). Target and distractor stimuli were distinguished by magnitude of distortion: 1 Bit/vertex for distractors, and 7.7 Bits/vertex for targets (Posner et al., 1965). All objects were rendered as coloured regions. The colours of all items were selected randomly with replacement from the default colour palette in R and were drawn at 50% opacity. Each item was positioned at a random, under the constraint that the object not extend beyond the bounds of an imaginary 6° x 6° box concentric with the centre of the display.

Stimulus images were generated in yoked target-absent/target-present pairs. The target-absent image within a pair contained only distractor objects. The yoked target-present image was identical, except that one distractor was replaced with a target, centred at the same position and drawn in the same colour. A total of 400 pairs of images was generated, and were sorted randomly into two sets, A and B, of 200 pairs each.

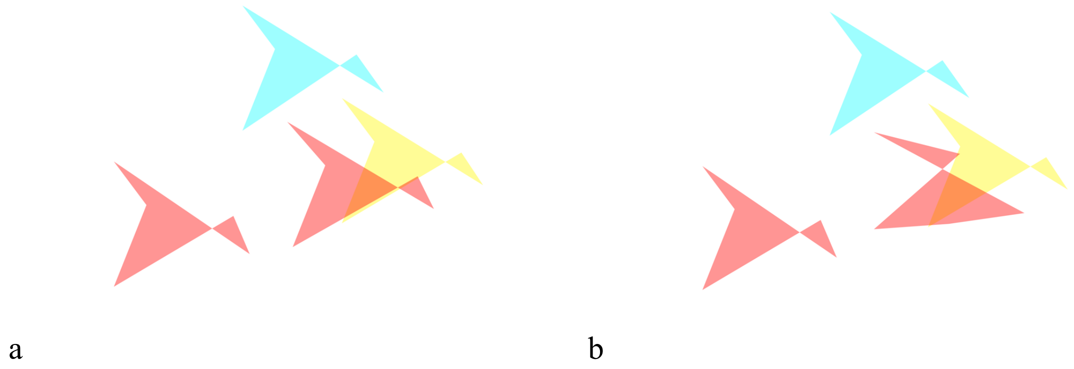

Figure 3.1. An example of a generated stimulus pair; a – target-absent, and b – target-present.

Procedure

Procedure was similar to that of Enright and McCarley (2018). Participants completed the visual search task in the same testing room in all conditions. In the single observer condition, participants worked independently, sitting at separate workstations at perpendicular angle to one another. Participants were instructed to refrain from communicating with each other and to look only at their own display. In the team condition, participants sat side-by-side at one workstation.

Instructions were presented onscreen at the start of the experimental session. They framed the search as a mock cell pathology screening task. The instructions informed participants that their task was to decide if a 'highly abnormal cell' was present (signal-plus-noise event) or not (noise-alone event) in each 'cell' sample. Each trial began with a fixation interval lasting 1000ms. The stimulus image and response rating scale were then presented for free viewing. Responses were made via mouse click on a six-point confidence scale including the judgments *Definitely Yes*, *Probably Yes*, *Guess Yes*, *Guess No*, *Probably No*, and *Definitely No*. A feedback message of 'You found a highly abnormal cell!', 'Good judgement', 'You missed a highly abnormal cell!', or 'False alarm' followed each hit, correct rejection, miss, and false alarm respectively. For target present trials, *Definitely Yes*, *Probably Yes*, and *Guess Yes* were treated as correct responses. Similarly, for target absent trials, *Definitely No*, *Probably No*, and *Guess No* were treated as correct responses.

Each team completed one block of 200 trials in the single observer condition and one block of 200 trials in the team condition. Each block included 100 target-present and 100 target-absent trials. Block order was counterbalanced across teams. Half of the teams used stimulus set A for single searcher conditions and set B for the team search conditions. The remaining teams used set B for the single searcher condition and set A for team search. Trial order was randomized within blocks and yoked across participants in the single observer condition.

Analyses

Data were analysed in RStudio (www.rstudio.com) using the Hierarchical Bayesian Analysis of Recognition Memory package (hbmem; Morey, Pratte, & Rouder, 2008; Pratte, Rouder, & Morey, 2009; Pratte & Rouder, 2012), which contains functions for fitting hierarchical versions of equal and unequal variance Gaussian signal detection models to confidence rating data. The model is fit with a Bayesian Markov chain Monte Carlo (MCMC) sampling procedure, using vague priors on model parameters. The model fitting procedure fits was run for 10,000 burn-in iterations and 50,000 iterations for analysis.

Three versions of the model were fit. The first (EV) assumed equal-variance signal-plus-noise and noise-alone distributions. The second (UV, fixed S^2) assumed that the variance of the signal-plus-noise distribution might differ from that of the noise-alone distribution, but that it was fixed across observers. The third (UV, free S^2) allowed the variance of the signal-plus-noise distribution to vary across observers, assuming an additive effect of log variance (Pratte & Rouder, 2010). Model fittings were compared using the deviance information criterion (DIC; lower values indicate better performance), which measure the quality of model fit, accounting for the number of functional model parameters (Spiegelhalter, Best, Carlin, & van der Linde, 2002). It is possible to fit two versions of each model, one in which item effects are included in the modelling, thereby accounting for them in estimates of collaborative performance, and one in which item effects are excluded in the modelling, reintroducing the correlations between team members' responses. Here, we limit reporting to models that did not include item effects.

Because two single observers were associated with each team, data were not amenable to a conventional paired-samples comparison of individual versus team search conditions. Instead, search condition was treated as a between-subject variable with 32 participants in the single-searcher condition and 16 participants in the team search condition. All data were initially fit to the UV, free S^2 model (1000 burn-in iterations, 10,000 analysis iterations) to check that team members and their associated team met the data inclusion criterion of a minimum of 60% accuracy. Any data, both individual and collaborative, that failed to meet the minimum accuracy for inclusion were excluded from further analysis.

$UW_{\rho=0}$ model predictions were generated using the hierarchical group mean parameter estimates of μ_n, μ_s, and σ_s for the individual search condition at each iteration of the MCMC process. Because the modeling provided one group-level estimate of each parameter—that is, it did not provide separate estimates for two different searchers within a team— this analysis assumed that the two searchers comprising a team were equally sensitive, making predictions for the UW model equivalent to those for OW model. The Mock UW model predictions were generated by first averaging the two searchers' responses on each trial of the individual condition, truncating the result to place the result on a 6-point scale, and finally submitting the ratings to the hbmem model.

The reported data below present the means and 95% Bayesian credible intervals (BCI) of the posterior distributions given by the hbmem model. Data were plotted in R using the *ggplot2* package v 2.2.2 (Wickham & Chang, 2016), including the geom_density function for plots of posterior distributions.

Results

All participants and teams in Experiment 1 met the minimum 60% accuracy inclusion score. The DIC values for Experiment 1 and 2 are presented in table 3.1, and show that the UV, free S^2

produced the best fit. The results of that model are therefore reported below. The MCMC chains for single observers, teams, and mock UW model predictions for Experiment 1 are shown in Figure 3.2, and by inspection, appear to have converged.

Table 3.1. DIC values for the EVSD, UV, S^2 fixed and UV, S^2 free for Experiments 1 & 2

	DIC values		
	EV	UV, S^2 fixed	UV, S^2 free
Experiment 1	26544.14	26542.55	26468.21
Experiment 2	26504.53	26507.81	26482.71

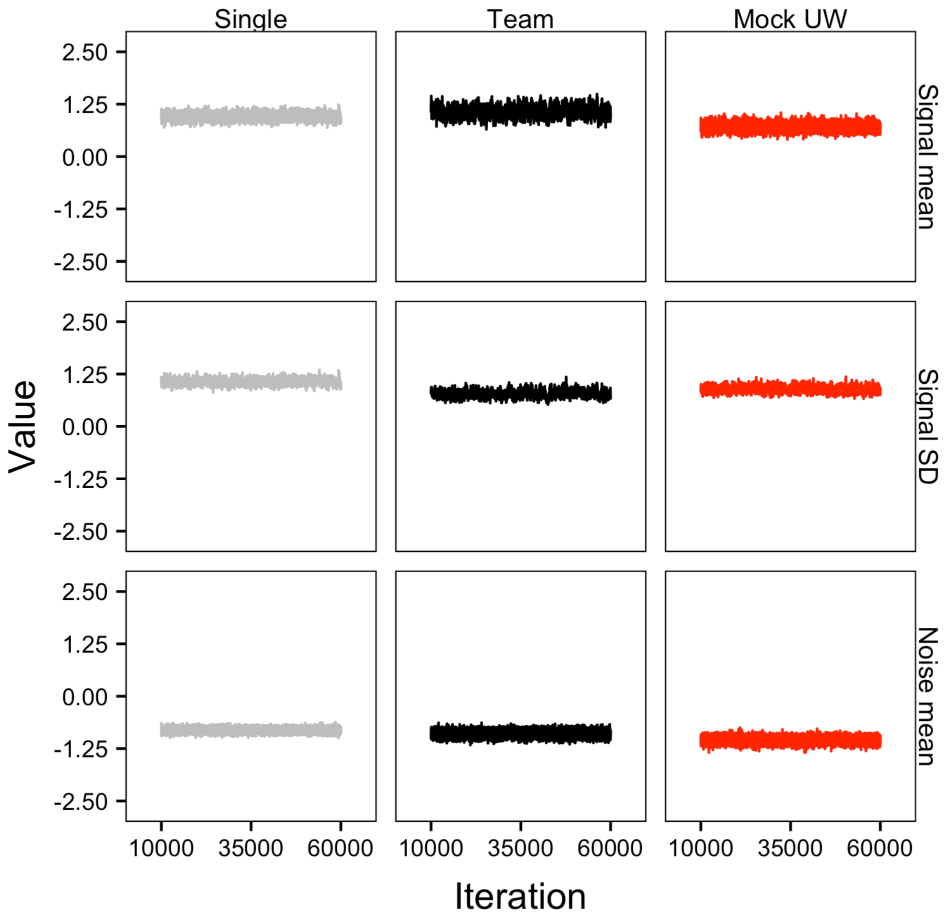

Figure 3.2. MCMC chains in Experiment 1. Rows show estimated parameters and columns the search condition.

The *z*ROCs for the single observers, teams, UW$_{\rho=0}$ model and mock UW model predictions, based on estimates of the group-level parameters, are presented in figure 3.3. The z-slopes for the signal and noise distributions were less than 1.0 (M = 0.81 for single observers and M = 0.87 for teams), indicating that the signal-plus-noise distribution had a larger variance than the noise-alone distribution.

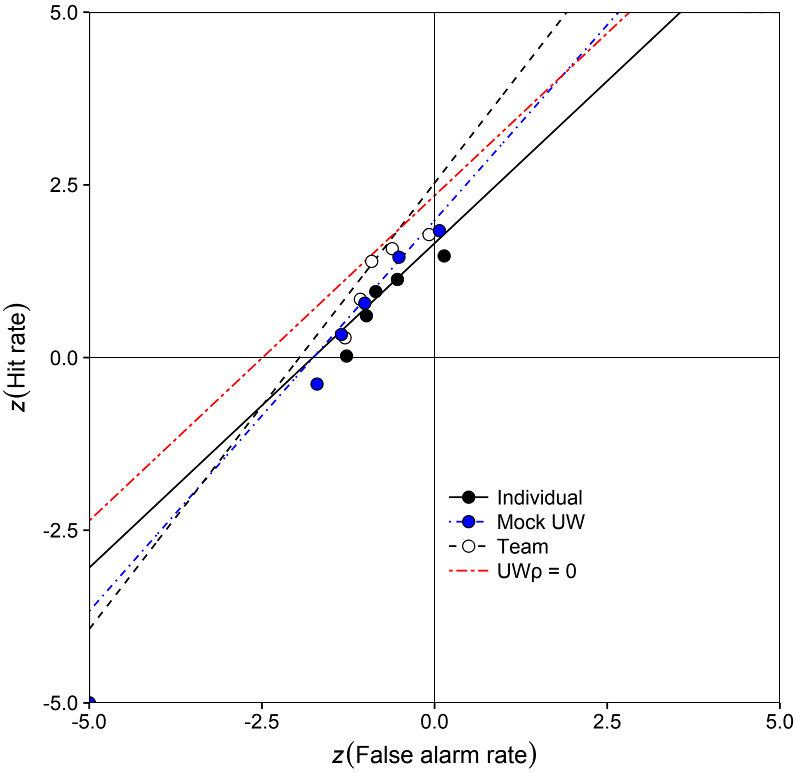

Figure 3.3. *z*ROCs for Experiment 1 with empirical data superimposed.

Figure 3.4 shows the posterior distribution of *d'$_e$* scores for single observers, teams, UW$_{\rho=0}$ model-predicted and mock UW model-predicted performance, based on estimates of the group-level parameters. Teams (M = 2.20, BCI[1.97, 2.43]) outperformed single observers (M = 1.71, BCI[1.57, 1.85], mean difference 0.49, BCI[0.22, 0.76]. Team performance fell between the mock UW model predictions (M = 1.87, BCI[1.66, 2.08]) and UW$_{\rho=0}$ model predictions (M = 2.41, BCI[2.22, 2.63]).

Figure 3.5 shows the distributions and 95% BCIs of the difference scores between observed and predicted team performance levels. The mean score for the mock UW model's predictions was negative, indicating that the model tended to underestimate observed sensitivity, with a BCI that just excluded a value of zero. The mean score for the the UW$_{\rho=0}$ model was positive, but with a BCI that overlapped zero.

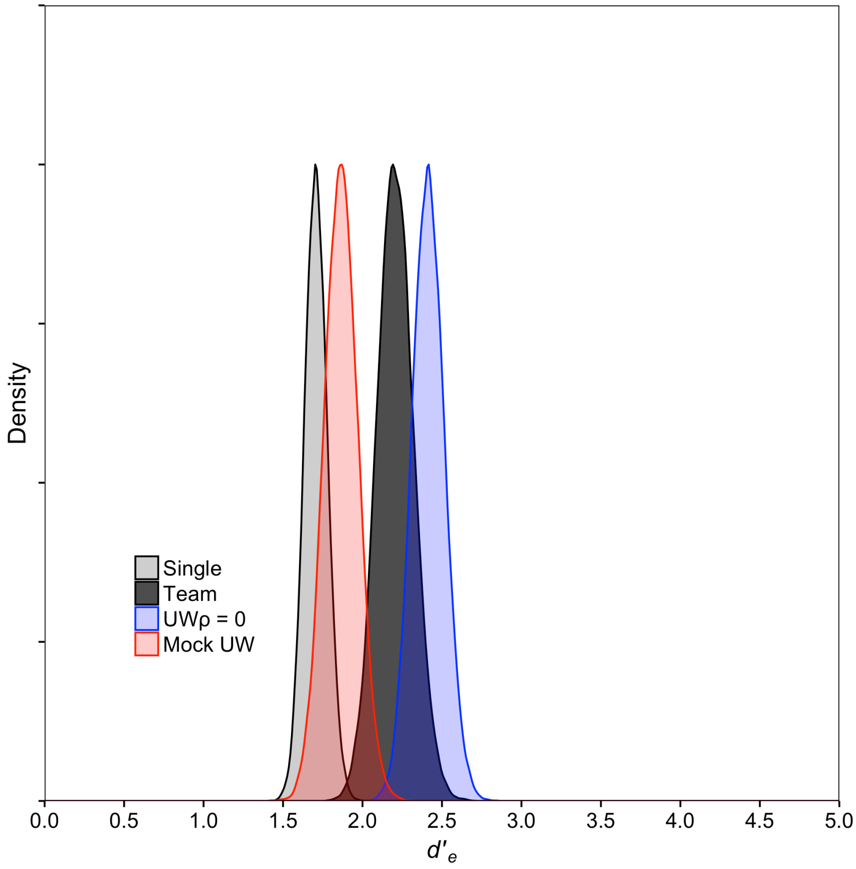

Figure 3.4. The posterior distributions of d'_e for single observers (light gray), team (dark gray), mock UW teams (red), and the UW$_{\rho=0}$ model (blue) in Experiment 1.

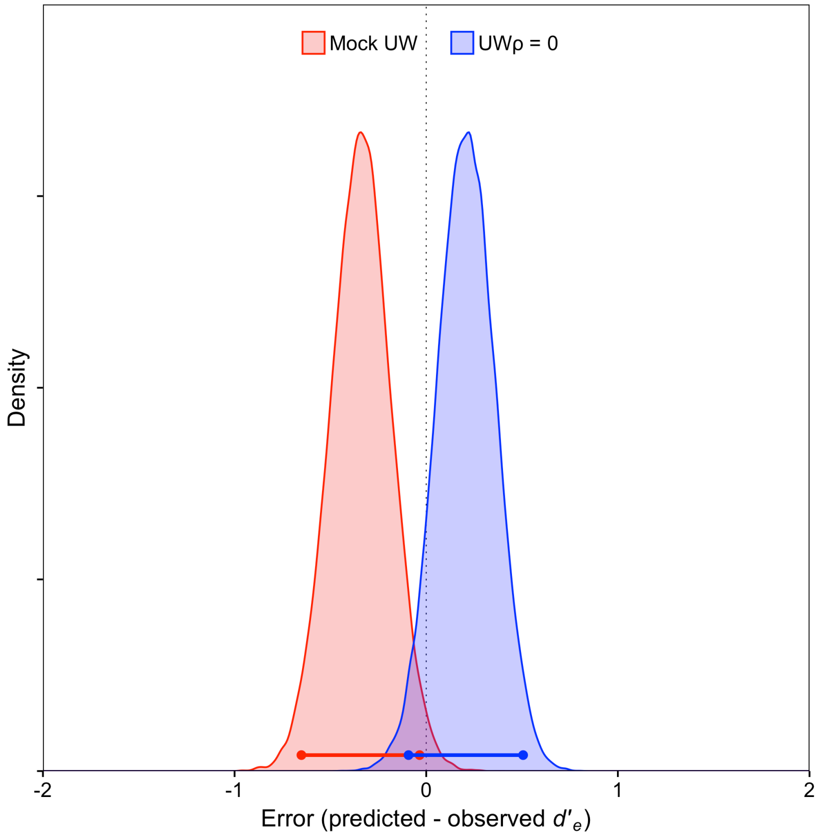

Figure 3.5. Distribution of difference scores between observed and predicted team performance in Experiment 1. Solid red and blue lines near the bottom of the figure indicate 95% BCIs.

A more specific analysis tested the fit of the mock UW team and $UW_{\rho=0}$ model predictions at the team-by-team level. Figure 3.6 shows the difference between observed and predicted team performance for all 16 teams. The mock UW model underestimated sensitivity for three teams (teams 1, 7, and 15) as their predictions differed credibly from the zero-error point of observed team performance. The $UW_{\rho=0}$ model overestimated sensitivity for two teams (teams 6, and 16).

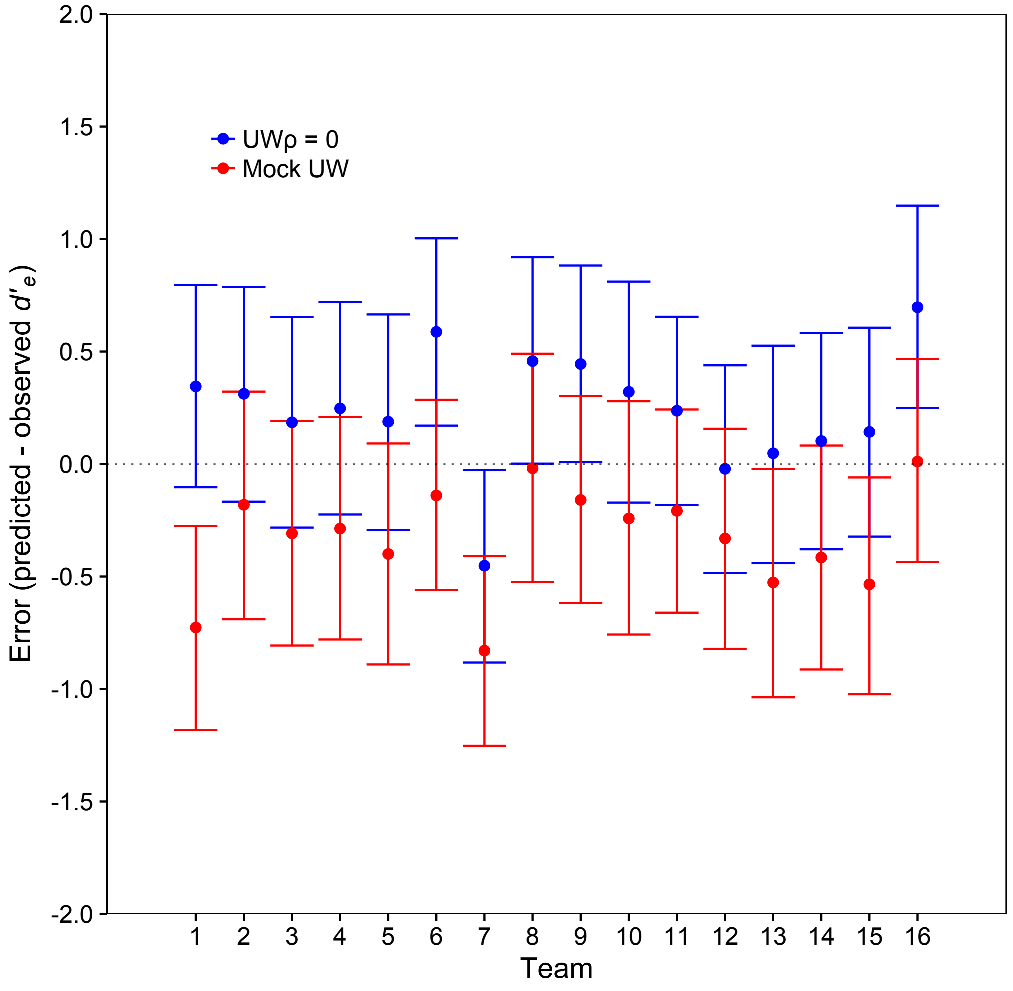

Figure 3.6. Difference between observed team sensitivity and UW$_{\rho=0}$ model (blue) and mock UW (red) sensitivity in Experiment 1. Error bars are 95% credible intervals on the difference between observed and predicted scores.

Discussion

Observed team sensitivity fell between the predictions of the mock UW and UW$_{\rho=0}$ models suggesting that teams performed above their expected performance levels given their correlated responses. This pattern mimics the results of Enright and McCarley (2018), who found that teams with correlated judgments also outperformed the predictions of a correlated UW model in a simulated baggage search task.

Enright and McCarley (2018) found that collaborative performance levels were similar regardless of whether team members were collocated. However, Yu and Wu (2015) found participants were quicker to detect targets in x-ray baggage images when searching in the presence of another. Liu and Yu (2017) also found participants produced shorter response times when detecting a “C” amongst “O” distractors when searching in the presence of another. Participants’ eye-movements reflected social facilitation effects such that fixations, saccades, and scan paths changed as a function of task complexity and search condition (e.g., alone versus in the presence of another; Liu & Yu, 2017). The social aspect in these studies was achieved using differing methods. Enright and McCarley (2018), for example, instructed collaborators to focus attention on their computer display, refrain from communicating and exit the testing room upon task completion. Team members, thus, may not have felt evaluated by the other team member. In the other studies (e.g., Liu & Yu, 2017; Yu & Wu, 2015), searchers were accompanied by an ‘examiner’ who focused attention on the searcher completing the task, leaving little space to avoid a sense of performance evaluation. Taken together, these findings suggest that it is possible that collaborative sensitivity is influenced by social facilitation effects. Experiment 2, thus, provided an opportunity to test whether Experiment 1’s findings generalize to non-collocated teams.

Experiment 2

Experiment 2 modified the procedure of Experiment 1 by asking participants to perform the collaborative search task from separate locations, communicating vocally. Experimental methods were preregistered (Enright, McCarley, & Leggett, 2018b, June 13) (https://osf.io/wcp69/?view_only=2069ee7afc434eef901820006a5f6665).

Participants

Thirty-two participants, making 16 pairs, of undergraduate students (23 female, M_{age} = 22.25, *SD* = 8.80) were recruited via Flinders University first-year research participant pool. All participants presented normal colour vision and normal or corrected-to-normal visual acuity and received $20AU in exchange for participation.

All stimuli and procedures were exactly the same as Experiment 1 except that participants completed both the individual and collaborative conditions in separate testing rooms. When performing the collaborative condition, teams communicated via Skype (www.skype.com) using only the phone function (i.e., no video), which operated on the same computer as the visual search task; however, no skype window was visible during the search task.

Results

All participants in Experiment 2 met the minimum 60% accuracy inclusion criteria. Figure 3.7 shows the MCMC chains for single observers, teams, and mock UW model predictions for Experiment 2. By inspection, the chains appear to have converged.

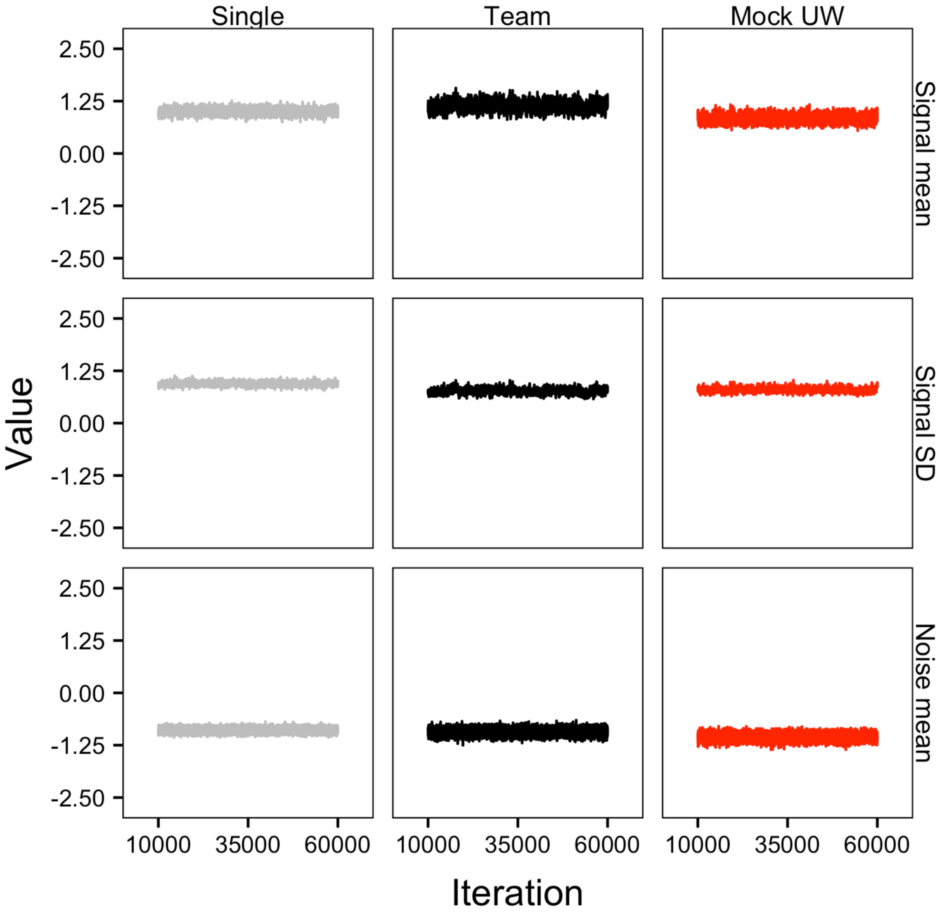

Figure 3.7. MCMC chains in Experiment 2. The rows show estimated parameters and the columns show search condition.

Figure 3.8 shows the *z*ROCs for the single observers, teams, $UW_{\rho=0}$ model and mock UW team model predictions, again based on estimates of the group-level parameters. The signal-plus-noise and noise-alone distributions' z-slopes were less than 1.0 (M = 0.83 for single observers and M = 0.80 for teams) indicating that the signal-plus-noise distribution had a larger variance than the noise-alone distribution.

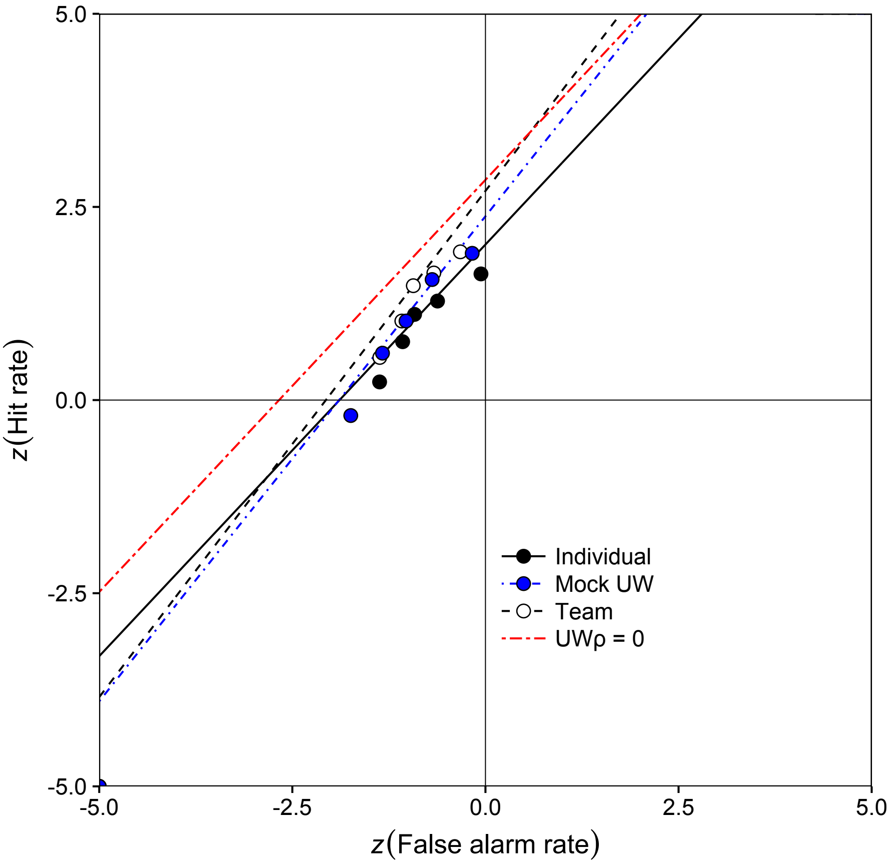

Figure 3.8. *z*ROCs for Experiment 2 with empirical data superimposed.

Figure 3.9 shows the posterior distribution of d'_e scores for single observers, teams, $UW_{\rho=0}$ model-predicted and mock UW model-predicted performance, based on estimates of the group-level parameters. Teams (M = 2.34, BCI[2.13, 2.56]) outperformed single observers (M = 1.95, BCI[1.81, 2.08]), mean difference (M = 0.40, BCI[0.14, 0.65]). Team performance was similar to mock UW teams (M = 2.11, BCI[1.91, 2.31]) and fell below $UW_{\rho=0}$ model (M = 2.76, BCI[2.57, 2.93]).

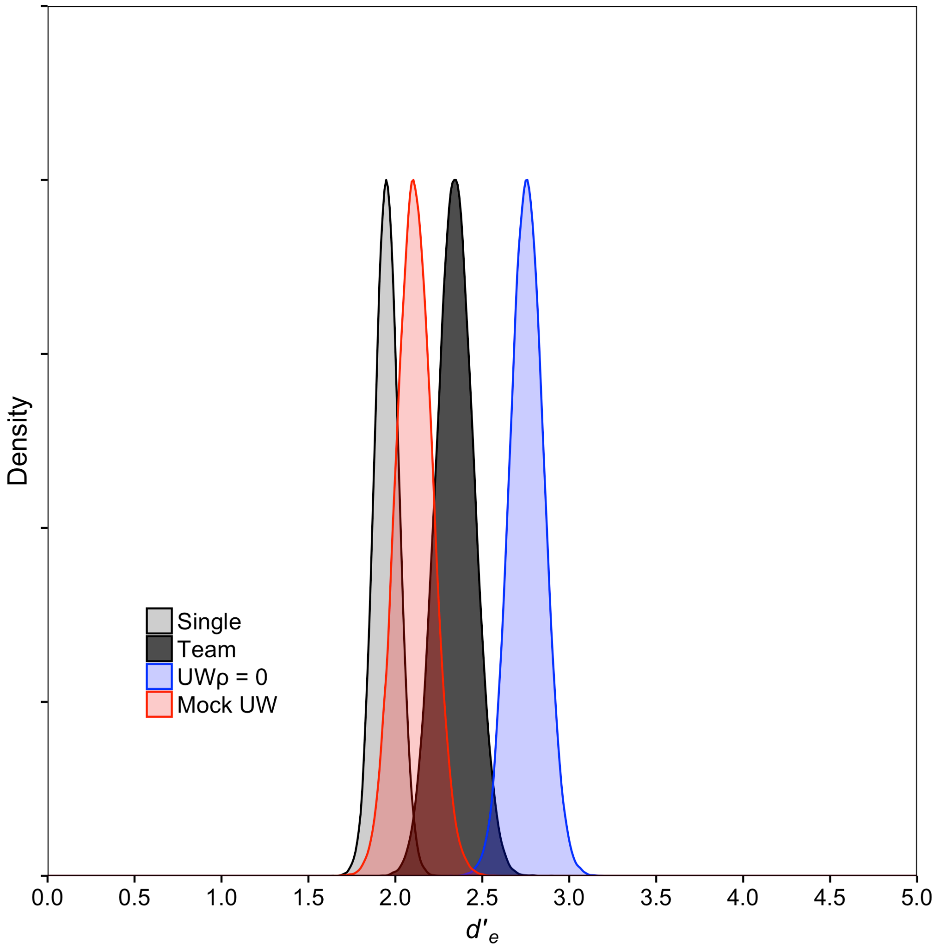

Figure 3.9. The posterior distributions of d'_e for single observers (light gray), team (dark gray), mock UW model-predicted (red), and $UW_{\rho=0}$ model-predicted (blue) performance in Experiment 2.

Figure 3.10 shows the distributions and 95% BCIs of the difference scores between observed and predicted team performance levels. Team performance was not credibly different from the mock UW model predictions and were credibly below the $UW_{\rho=0}$ model predictions.

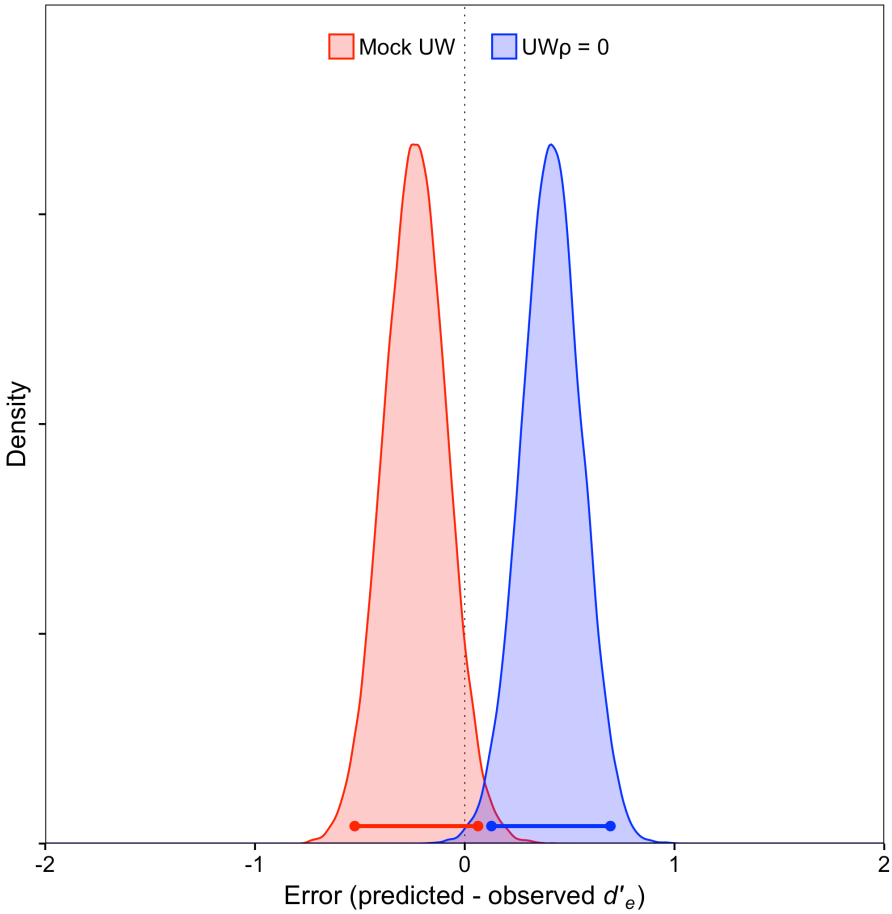

Figure 3.10. Distribution of difference scores between observed and predicted team performance in Experiment 2. Solid red and blue lines near the bottom indicate 95% BCIs.

A team-by-team level analysis tested the fit of the mock UW team and $UW_{\rho=0}$ model predictions (Figure 3.11). The $UW_{\rho=0}$ model overestimated sensitivity for seven teams (teams 5, 8, 9, 11, 13, and 14). The mock UW model underestimated sensitivity for one team (15).

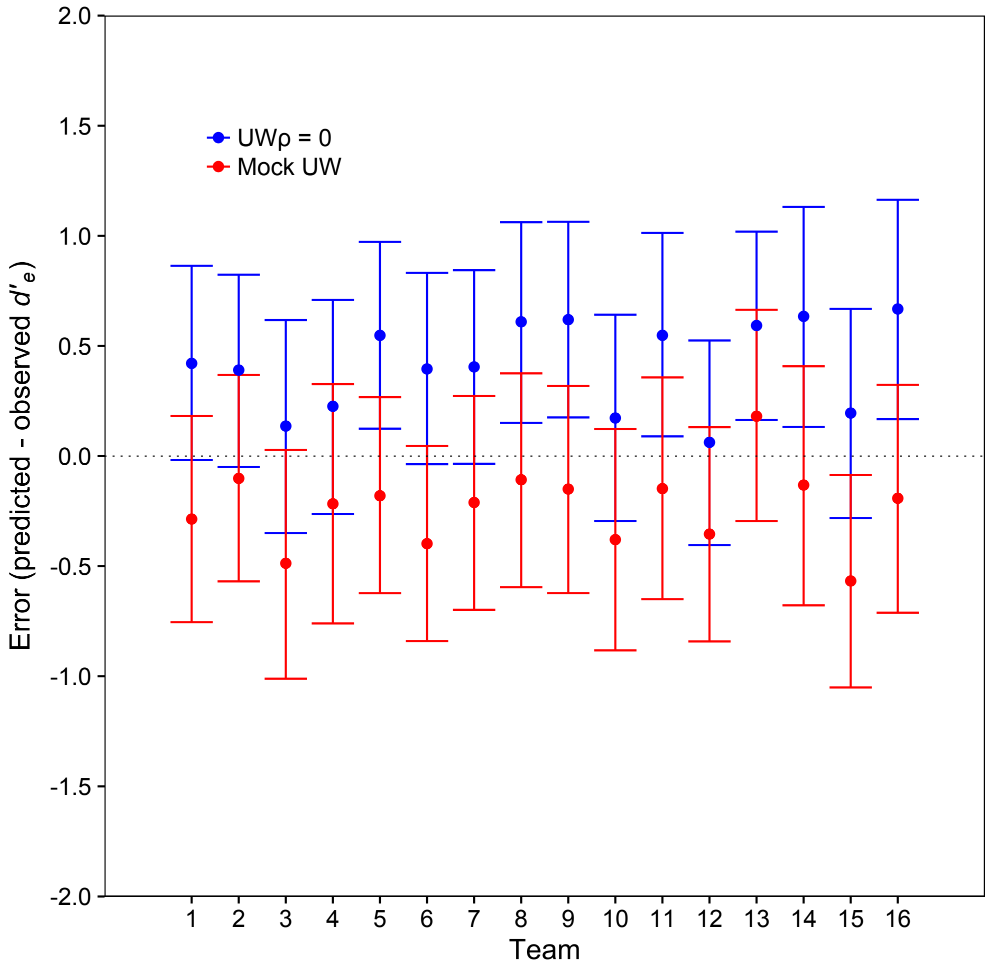

Figure 3.11. Difference between observed team performances and $UW_{\rho=0}$ model-predicted (blue) and Mock UW model-predicted performances (red) in Experiment 2. Error bars are 95% credible intervals on the difference between observed and predicted scores.

General Discussion

Collaborative performance in a visual search task was explored in two experiments. Observed team performance was compared to the predictions of two versions of a uniform weighting model – one that allowed stochastic dependency between team members' judgments and one that assumed stochastic independency. In Experiment 1, single observers performed the task in the same testing room whereas in Experiment 2, single observers performed the task in separate testing rooms. In both experiments, teams performed better than single observers. Collaborative sensitivity fell roughly midway between the predictions of both models and, more specifically, was slightly closer to the predictions of the uncorrelated UW model in Experiment 1 and the correlated UW model in Experiment 2.

Multiple previous studies found collaborative sensitivity levels above those of individuals and similar to those predicted by the UW model (e.g., Bahrami et al., 2010; 2012; Hinsz, 1990; Sorkin et al., 2001). Our methods are unique in that we expected correlated judgments and unequal variances in the signal-plus-noise and noise-alone distributions

Our results replicate Enright and McCarley's (2018) findings and extend them to a signal detection task that limits top-down knowledge of the target. Participants were aware that targets would appear distorted from other stimulus items, but were unaware of what colour, shape, size, or spatial orientation the target might present. Despite this lack of target information, team performance levels were above those from single observers. Visual search studies that constrain top-down target guidance show that search is generally less efficient (Chen & Zelinsky, 2006; Schmidt & Zelinsky, 2009). Our results, however, suggest collaboration during visual search tasks that include limited top-down knowledge of the target is valuable.

Similar to Enright and McCarley (2018), collaborative performance was better than expected given team members' correlated responses. It seems improbable that team members engaged a divide-and-conquer approach reflecting the effects of both dividing the search display and adopting differing target templates, though (Chen, 2007; Malcolmson et al., 2007). Chen (2007) found that pairs of observers divided the search display to share the search labour. Similarly, Malcolmson et al. (2007) found teams reported using the same strategy. Dividing the stimulus display in this way seems unlikely in the current task because participants needed to compare a potential target with at least two other items in the search display. Similarly, it is unlikely that team members adopted differing target templates because top-down knowledge of targets was constrained.

It is possible, though, that team performance was greater than what was expected because team members may have put forth more effort when working collaboratively, to avoid negative performance evaluation (Kerr & Tindale, 2004). Team members may have engaged in social comparison, thereby increasing their individual sensitivity levels before integrating judgments to reach a joint decision. This explanation fits well with our findings because team performance was credibly below the correlated UW model predictions, and no longer above performance expectations, when teams collaborated from separate rooms (Experiment 2). Although collaborative sensitivity was slightly lower for non-collocated teams, the different patterns of observed collaborative performance in Experiment 1 and 2 are potentially no more than random variation, suggesting that team collocation, and consequently social facilitation effects, are unlikely contributing to obtained collaborative benefit.

Future research, with larger sample sizes or using a within-subject comparison, is required to determine whether collaborative sensitivity differs for collocated and non-collocated conditions. The exact search strategies teams employed remains unclear and so future research would benefit from explicitly measuring which strategies teams use.

Constraints on generality

Generalizing the current findings is limited to the characteristics of the sample population and the task (Simons, Shoda, & Lindsay, 2017). The above findings were produced by non-expert screeners; expert screeners might produce different results due to potentially different scan paths (Nodine & Kundel, 1987). Participants performed the visual search task in a quiet room with no distractions. Future research is needed to confirm if our obtained pattern of results generalise to less controlled conditions.

This concludes the currently available paper.

CHAPTER 4:

GENERAL DISCUSSION

A broad aim of this thesis, thus, was to better understand collaborative visual search. More specifically, this thesis had two main aims: 1) to replicate previous findings (i.e., collaborative search performance that matches or exceeds the predictions of the uniform weighting model) and extend them to a signal detection task using naturalistic stimuli, and 2) to investigate a joint search strategy when little target information is available to observers.

Study 1 aimed to replicate previous collaborative visual search findings, that is collaborative search performance on par with or better than the predictions of a UW model (e.g., Bahrami et al., 2010, 2012; Malcolmson et al., 2007), and extend findings to a signal detection task using naturalistic stimuli. Benchmarking collaborative search performance relative to the predictions of the UW model in our signal detection task required addressing two obstacles: the correlation between collaborators' judgments and signal-plus-noise and noise-alone distributions with differing variance.

The correlation between collaborators' judgments is problematic because the standard equation for predicting UW-performance assumes either that team members provide stochastically independent judgments, or that the correlation between their judgments is known (Sorkin et al., 2001). Collaborative searchers in real-world contexts are likely to provide stochastically dependent judgments, though, because searchers are jointly inspecting the same stimulus thereby increasing the shared variance in their information encoding. When team members share a large proportion of shared variance, they will contribute redundant information that increases the similarity between their judgments and consequently decreases the value of collaboration (Sorkin & Dai, 1994; Sorkin et al., 2001).

Differing variance in the signal-plus-noise and noise-alone distributions requires addressing because the UW model predictions were originally formulated in terms of *d'*, which assumes equal-variance evidence distributions. Naturalistic stimuli will often violate this equal-variance assumption, because the signal-plus-noise distribution will include the variance associated with the signal and the noise, whereas the noise distribution contains the variance associated with only the noise (Swets, 1986).

Both problems (i.e., correlated team member responses and unequal variance in the evidence distributions) were circumvented in Study 1 by analysing receiver operating characteristics (ROC; Green & Swets, 1966; Macmillan & Creelman, 2005). The UW model was adapted to predict collaborative signal detection performance in ROC space. The correlated UW model predictions were derived by averaging team members' trial-by-trial confidence ratings, an analysis that incorporates the correlations between observers' judgments. Data were fit to Bayesian hierarchical models that produced posterior distributions of d'_e, a sensitivity index that does not require the same variance in the signal-plus-noise and noise-alone distributions.

To benchmark team sensitivity levels relative to individual sensitivity levels and gauge collaborative performance relative to both versions of the UW model predictions, participants performed a simulated baggage screening task, working independently or collaboratively in two-person teams. In Experiment 1, participants completed the individual search condition in separate testing rooms and the collaborative search condition in the same testing room, sharing one computer. In Experiment 2, they completed both the individual and collaborative search conditions in the same testing room. In Experiment 3, they completed both the individual and collaborative search conditions in separate testing rooms. Finally, in Experiment 4, they completed both search conditions in the same testing room, and stimulus presentation time was fixed to 3s.

Results confirmed that teams outperformed single searchers in all four experiments. Somewhat surprisingly, though, team sensitivity levels fell mostly in between the predictions of the correlated and uncorrelated UW model predictions with some modest disagreement in the pattern of statistically credible differences across the four experiments. Although the pattern of credible differences fluctuated across experiments, team sensitivity fell between the predictions of the uncorrelated and correlated UW model predictions.

The pattern of results for Experiments 1-4 likely reflected statistical variability and not real differences in collaborative performance levels. A meta-analysis of the data produced in the four experiments tested this statistical variability assumption and provided a more accurate picture of the team performance levels relative to the UW model's predictions. Results of the meta-analysis showed that team sensitivity fell midway between the predictions of both versions of the UW model and, more specifically, were credibly below the predictions of the correlated UW model and credibly above the correlated UW model predictions.

We considered three potential explanations for why teams performed better than expected given the statistical limitations. The first possibility is that speed-accuracy tradeoff effects (Reed, 1973; Wickelgren, 1977) mimicked performance levels above the correlated UW model's predictions. A speed-accuracy tradeoff effect was hypothesized because teams in Experiments 1-3 took significantly longer to respond than did individuals, indicating the possibility that teams employed longer scanning durations thereby increasing accuracy. However, Experiment 4 tested this assumption by limiting stimuli presentation time to 3s, and results showed that collaborative performance remained above the correlated UW model predictions. This suggested speed-accuracy tradeoff effects unlikely contributed to the collaborative performance levels.

The second potential account for why teams outperformed the correlated UW model predictions is that individuals may have put forth greater effort when collaborating to avoid negative performance evaluation. Such social facilitation effects (Kerr & Tindale, 2004) might have acted to effectively increase individuals' sensitivity levels when working collaboratively, thereby producing higher individual sensitivity levels before aggregating responses to reach a joint decision.

The third possibility is that teams might have interacted during collaboration in a way that changed their information sampling, effectively increasing the individual *d'* scores that contribute to UW performance, or decorrelating their individual judgments. A division-of-labour strategy, for example, reflects two effects. One, team members search particular areas of the search display, dividing the search labour (Chen, 2007; Malcolmson et al., 2007). And two, paired searchers may have also applied different top-down target templates for the different potential knife target items, guiding attention to the target more easily. Doing so could have reduced the similarity in their encoded evidence for the signal, thereby decorrelating their individual judgments before reaching a joint decision.

Study 2 aimed to first replicate Study 1's findings using different stimuli, and second, limit participants' capacity to engage a divide-and-conquer search strategy by limiting target certainty. Study 2 reduced top-down target information using dot-distortion stimuli (Posner, Goldsmith, & Welton, 1967; Smith, Redford, Gent, & Washburn, 2005). Stimuli were unique each trial, limiting observers' target information and capacity to use attentional guiding to detect targets. Furthermore, because the target was defined as an extreme distortion among more modest distortions—that is, an outlier in a space of shapes—the task did not allow participants to identify the target based strictly on its own characteristics but required them to compare presented stimulus items. This limited the option to divide the search display between observers. In Experiment 1, teams performed the individual search condition in the same testing room using separate testing stations, and the collaborative condition in the same room, sharing one computer. In Experiment 2, participants completed both the individual search condition and the collaborative search condition in separate testing rooms. Data were again analysed using Bayesian hierarchical analysis of receiver operating characteristic curves.

Results replicated the effects of Study 1. In Experiment 1, team performance levels were slightly above the predictions of the uncorrelated UW model predictions and not credibly different from the predictions of the correlated UW model. In Experiment 2, team performance was credibly poorer than the uncorrelated UW model predictions and not credibly different from the correlated UW model predictions. Overall, team sensitivity fell roughly in between the predictions of the correlated and uncorrelated UW models. Study 2 thus extends these findings to a signal detection task that assumed observers had limited top-down target knowledge.

Given that top-down target information was limited, it seems unlikely that a division-of-labour strategy contributed to the collaborative benefit obtained. Although non-collocated teams performed slightly poorer than collocated teams, this pattern of results is likely to largely reflect statistical variation, rather than a true, substantial difference between performance levels. It also seems somewhat unlikely that social facilitation effects explain teams' performance levels because collaborative sensitivity was similar regardless of whether teams were collocated.

Practical implications and future directions

Our findings carry implications most specifically for the human factors of fields in which collaborative visual search is common, e.g., transportation security and medical image reading. Our most consistent finding is that collaborative searchers outperform single searchers. This suggests that agents working to detect a signal are more efficient when searching collaboratively than independently.

In fact, some of our findings indicate that teams can even outperform what is expected given their individual sensitivity levels and the similarity in team members' judgments. Such findings imply that teams can adopt visual search performance strategies that work to decorrelate their judgments to result in a larger collaborative benefit when integrating their judgments. The exact method of how participants achieve de-correlated judgments, though, is still unclear. As such, future research would benefit from directly investigating the visual search strategies teams employ, that increase their collaborative benefit, potentially using eye-tracking or dialogue analysis.

Teams continued to outperform individuals despite reduced target information, limiting teams' capacity to engage a divide-and-conquer search strategy. Previous studies (e.g., Chen & Zelinsky, 2006; Schmidt & Zelinsky, 2009; Vickery, King, & Jiang, 2005) found visual search performance is less efficient when top-down target knowledge is constrained. Our results show that collaboration is nonetheless valuable in such conditions. This finding carries implications in real-world contexts such as transportation security screening, in which officers are required to search for illicit items that take various shapes, sizes, and orientations. Under such conditions, it seems reasonable to suggest collaborative searchers will detect targets with greater sensitivity than individual searchers.

Results also demonstrate that non-collocated teams can perform similarly to collocated teams. Remotely collaborating teams produced similar performance levels suggesting that collaborative gains do not require team members to work in a face-to-face context to gain collaborative benefit. However, further research with larger sample sizes and a within-subject design is needed to confidently claim team collocation is not an important factor to collaborative visual search.

It is worth noting that our data might also be compatible with other models of visual data, such as Yonelinas' dual-processing model (Yonelinas, Dobbins, Szymanski, Dhaliwal, & King, 1996; Pratte & Rouder, 2011). Yonelinas' dual-processing model (Yonelinas, Dobbins, Szymanski, Dhaliwal, & King, 1996; Pratte & Rouder, 2011) is effectively a signal detection model with the probability of a discrete detection state. Dual process models essentially argue that target detection can occur due to threshold models, discrete detection, or signal detection models, familiarity compared to a criterion. Some of the modelled data presented here (Study 1) show very high d'_e and σ values. These exceptionally high values might suggest a discrete detection, rather than a signal detection, process. Future work comparing discrete state and signal detection models would be useful.

The high σ values obtained in the modelled data presented in Study 1 drop considerably in Study 2. In fact, σ is roughly halved when participants' task limits precise target information. One possible explanation is that Study 2 included targets that were much more difficult to detect with certainty, resulting in participants relying on less extreme values when responding, indicating less confidence in their judgements.

Constraints on generality

As Simons, Shoda, and Lindsay (2017) note, generalising the above findings is limited to the context of the experimental tasks and the characteristics of the sample population. Participants performed our visual search tasks in quiet rooms with no distractions. Further research will be needed to confirm that effects replicate under in real-world conditions. We also sampled a population of non-expert observers. Future research, thus, is needed to confirm if our results generalise to expert observers.

Generalizing our obtained pattern of results (Studies 1 and 2) is also limited to events with higher than some naturally occurring target prevalence rates. Our studies included signal rates of 40-50% whereas real-world contexts reported signal rates closer to 2% (Wolfe et al., 2007). Target prevalence aligned with naturalistic environments could trigger a low-prevalence effect in which observers more often miss targets because their presentation is less likely (Wolfe et al., 2007). Additional work will be necessary to confirm that the current effects persist under conditions of extremely low target prevalence.

REFERENCES

Bahrami, B., Olsen, K., Bang, D., Roepstorff, A., Rees, G., & Frith, C. D. (2012). What failure in collective decision-making tells us about metacognition. *Philosophical Transactions of The Royal Society B, 367*, 1350-1365. doi: 10.1098/rstb.2011.0420

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally Interacting Minds. *Science, 329*, 1081-1085. doi: 10.11.26/science.1185718

Baltes, B. B., Dickson, M. W., Sherman, M. P., Bauer, C. C., & LaGanke, J. S. (2002). Computer-Mediated Communication and Group Decision Making: A Meta-Analysis. *Organizational Behavior and Human Decision Processes, 87*, 156-179. doi: 10.1006/obhd.2001.2961

Bang, D., Fusaroli, R., Tylén, K., Olsen, K., Latham, P. E., Lau, J. Y. F., Roepstorff, A., Rees, G., Frith, C. D., & Bahrami, B. (2014). Does interaction matter? Testing whether a confidence heuristic can replace interaction in collective decision-making. *Consciousness and Cognition, 26*, 13-23. doi: 10.1016/j.concog.2014.02.002.

Beck, M. R., Lohrenz, M. C., & Trafton, J. G. (2010). Measuring search efficiency in complex visual search tasks: Global and local clutter. *Journal of Experimental Psychology Applied*, *16*, 238–250. http://doi.org/10.1037/a0019633

Benbasat, I., & Lim, L-H. (1993). The Effects of Group, Task, Context, and Technology Variables on the Usefulness of Group Support Systems: A meta-analysis of experimental studies. *Small Group Research, 24*, 430-462. doi: 10.1177/1046496493244002

Brennan, S. E., Chen, X., Dickinson, C. A., Neider, M. B., & Zelinsky, G. J. (2008). Coordinating cognition: The costs and benefits of shared gaze during collaborative search. *Cognition*, *106*(3), 1465–1477. https://doi.org/10.1016/j.cognition.2007.05.012

Burgess, A. E., & Colborne, B. (1988). Visual signal detection. IV. Observer inconsistency. *JOSA A, 5*, 617-627.

Chen, X. (2007). State University of New York at Stony Brook, ProQuest Dissertations Publishing, 3337503.

Chen, X., & Zelinsky, G. J. (2006). Real-world visual search is dominated by top-down guidance. *Vision Research, 46*, 4118-4133. doi: 10.1016/j.visres.2006.08.008

Chidambaram, L., & Tung, L. L. (2005). Is Out of Sight, Out of Mind? An Empirical Study of Social Loafing in Technology-Supported Groups. *Information systems research, 16*, 149-168. doi: 10.1287/isre.1050.0051

Condorcet, M. D. (1785). Essay on the Application of Analysis to the Probability of Majority Decisions. *Paris: Imprimerie Royale.*

Davis, J. H. (1992). Some compelling intuitions about group consensus decisions, theoretical and empirical research, and interpersonal aggregation phenomena: Selected examples 1950-1990. *Organizational Behavior and Human Decision Processes, 52*, 3-38. doi: 10.1016/0749-5978(92)90044-8

Davis, J. H. (1996). Group decision making and quantitative judgments: a consensus model. In *Understanding Group Behaviour: Consensual Action by Small Groups*, ed. E Witte, JH Davis, 1:35-59. Mahwah, NJ: Erlbaum.

Dee, F. R. (2009). Virtual microscopy in pathology education. *Human Pathology; Philadelphia, 40*, 1112-21. doi: 10.1016/j.humpath.2009.04.010

Dee, F. R., Lehman, J. M., Consoer, D., Leaven, T., & Cohen, M. B. (2003). Implementation of virtual microscope slides in the annual pathobiology of cancer workshop laboratory. *Human Pathology; Philadelphia, 34*, 430-436. doi: 10.1016/S0046-8177(03)00185-0

Denkiewicz, M., Raczaszek-Leonardi, J., Migdal, P., & Plewczynski, D. (2013). Information-Sharing in Three Interacting Minds Solving a Simple Perceptual Task. *Proceedings of the Annual Meeting of the Cognitive Science Society, 35.* https://escholarship.org/uc/item/2fj353sr.

Deutsch, M. (1958). Trust and Suspicion. *The Journal of Conflict Resolution, 2*, 265-279. http://www.jstor.org/stable/172886

Duncan, J., & Humphreys, G. W. (1989). Visual Search and Stimulus Similarity, *Psychological Review, 96,* 433-458. doi: 10.1037//0033-295X.96.3.433

Eckstein, M. P. (1998). The lower efficiency for conjunctions is due to noise and not serial attentional processing. *Psychological Science, 2*, 111-118. doi: 10.1111/1467-9280.00020

Eckstein, M. P., Thomas, J. P., Palmer, J., & Shimozaki, S. S. (2000). A signal detection model predicts the effects of set size on visual search accuracy for feature, conjunction, triple conjunction, and disjunction displays. *Perception & Psychophysics, 62*, 425-451. doi: 10.3758/bf03212096

Eidels, A., Houpt, J. W., Altieri, N., Pei, L., & Townsend, J. T. (2011). Nice guys finish fast and bad guys finish last: Facilitatory vs. inhibitory interaction in parallel systems. *Journal of Mathematical Psychology*, *55*, 176–190. https://doi.org/10.1016/j.jmp.2010.11.003

Egan, J., Schulman, A. I., & Greenberg, G. Z. (1959). Operating Characteristics Determined by Binary Decisions and by Ratings. *The Journal of the Acoustical Society of America, 31*, 768-773. doi: 10.1121/1.1907783

Enright, A., & McCarley, J. S. (2018, May 16). Collaborative Search in a Mock Baggage Screening Task: A Bayesian Hierarchical Analysis. https://doi.org/10.17605/OSF.IO/8975X

Enright, A., & McCarley, J. S. (2018, June 21). Collaborative searchers outperform individuals in the absence of precise target information. https://doi.org/10.17605/OSF.IO/V6CQK

Erev, I., Bornstein, G., & Galili, R. (1993). Constructive intragroup competition as a solution to the free rider problem: a field experiment. *Journal of Experimental Social Psychology, 29*, 463-478. doi: 10.1006/jesp.1993.1021

Forsyth, D. R. (1998). Methodological Advances in the Study of Group Dynamics. *Group Dynamics: Theory, Research, and Practice, 2*, 211-212. doi: 1089-2699/98/$3.00

Galton, F. (1907). The Wisdom of Crowds. *Nature, 75,* 450-451.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* New York: Wiley.

Hastie, R., & Kameda, T. (2005). The robust beauty of majority rules in group decisions. *Psychological review, 112*, 494. doi: 10.1037/0033-295x.112.2.494

Hill, G. W. (1982). Group Versus Individual Performance: Are *N* + 1 Heads Better Than One? *Psychological Bulletin, 91*, 517-539. doi: 0033-2909/82/9103-0517$00.75

Hinsz, V. B. (1990). Cognitive and consensus processes in group recognition memory performance. *Journal of Personality and Social Psychology, 59*, 705.

Hout, M. C. (2013). Target "templates": How the precision of mental representations affects attentional guidance and decision-making in visual search. *Arizona State University, ProQuest Dissertations Publishing.*

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research, 40*, 1489-1506. doi: 10.1016/S0042-6989(99)00163-7

Kameda, T., Tsukasaki, T., Hastie, R., & Berg, N. (2011). Democracy under uncertainty: The wisdom of crowds and the free-rider problem in group decision making. *Psychological Review, 118,* 76-96. doi: 10.1037/a0020699

Karau, S. J., & Williams, K. D. (1993). Social loafing: a meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology, 65*, 681-706. doi: 10.1037//0022-3514.65.4.681

Karau, S. J., & Williams, K. D. (1997). The effects of group cohesiveness on social loafing and social compensation. *Group Dynamics: Theory, Research, and Practice, 1*, 156-168. doi: 10.1037//1089-2699.1.2.156

Kerr, N. L., & Kaufman-Gilliland, C. M. (1994). Communication, commitment, and cooperation in social dilemma. *Journal of Personality and Social Psychology, 66*, 513-529. doi: 10.1037/0022-3514.66.3.513.

Kerr, N. L., & Tindale, R. S. (2004). Group performance and decision making. *Annual Review Psychology, 55*, 623-655. doi: 10.1146/annurev.psych.55.090902.142009

Kiesler, S., & Cummings, J. N. (2002). What do we know about proximity and distance in work groups? A legacy of research. *Distributed work*, *1*, 57-80.

Köhler, O. (1926). Kraftleistungen bei Einzelund Gruppenabeit [Physical performance in individual and group situations]. *Ind.Psychotech, 4*, 209-226.

Kruschke, J. (2015). Bayesian estimation of correlations and differences of correlations with a multivariate normal. Retrieved from http://doingbayesiandataanalysis.blogspot.com/2017/06/bayesian-estimation-of-correlations-and.html#comment-form.

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

Latané, B., Williams, K., & Harkins, S. (1979). Many hands make light the work: the causes and consequences of social loafing. *Journal of Personality and Social Psychology, 37*, 822-832. doi: 10.1037//0022-3514.37.6.822

Laughin, P. R., Bonner, B. L., & Miner, A. G. (2002). Groups perform better than the best individuals on letters-to-numbers problems. *Organizational Behaviour and Human Decision Processes, 88,* 605-620. doi: 10.1016/s0749-5978(02)00003-1

Levine, J. M., & Moreland, R. L. (1990). Progress in small group research. *Annual Review of Psychology, 41*, 585-634. doi: 10.1146/annurev.ps.41.020190.003101

Liu, N., & Yu, R. (2017). Influence of social presence on eye movements in visual search tasks. *Ergonomics, 60*, 1667-1681. doi: 10.1080/00140139.2017.1342870

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of the crowd effect. *Proceedings of the National Academy of Sciences of the United States of America, 108*, 9020-9025. doi: 10.1037/pnas.1008636108

Macmillan, N. A., & Creelman, C. D. (2005). *Detection Theory: A User's Guide*. Lawrence Erlbaum Associates: New York.

Malcolm, G. L., & Henderson, J. M. (2009). The effects of target template specificity on visual search in real-world scenes: Evidence from eye movements. *Journal of Vision, 9*, 1-13. doi: 10.1167/9.11.8

Malcolmson, K. A., Reynolds, M. G., & Smilek, D. (2007). Collaboration during visual search. *Psychonomic Bulletin & Review, 14*, 704-709. doi: 10.3758/bf03196825

McCarley, J. S. (2009). Effects of speed-accuracy instructions on oculomotor scanning and target recognition in a simulated baggage x-ray screening task. *Ergonomics, 52*, 325-333. doi: 10.1080/00140130802376059

McCarley, J. S., Kramer, A. F., Wickens, C. D., Vidoni, E. D., & Boot, W. R. (2004). Visual skills in airport-security screening. *Psychological Science, 15,* 302-306.

Mello-Thoms, C. (2006). The problem of image interpretation in mammography: Effects of lesion conspicuity on the visual search strategy of radiologists. *British Journal of Radiology*, *79*, S111–S116.

Menneer, T., Barrett, D. J. K., Phillips, L., Donnelly, N., & Cave, K. R. (2007). Costs in searching for two targets: Dividing search across target types could improve airport security screening. *Applied Cognitive Psy- chology, 21,* 915–932. doi:10.1002/acp.1305

Menneer, T., Cave, K. R., & Donnelly, N. (2009). The cost of searching for multiple targets: Effects of practice and target similarity. *Journal of Experimental Psychology: Applied, 15,* 125–139. doi:10.1037/a0015331

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*, 60-75.

Moon, J. Y., & Sproull, L. (2001). Turning Love into Money: How some firms may profit from voluntary electronic customer communities.

Mordkoff, J. T., & Yantis, S. (1991). An interactive race model of divided attention. *Journal of Experimental Psychology: Human Perception and Performance*, *17*, 520.

Morey, R. D., Pratte, M. S., & Rouder, J. N. (2008). Problematic effects of aggregation in *z*ROC analysis and a hierarchical modeling solution. *Journal of Mathematical Psychology, 52*, 376-388. doi: 10.1016/j.jmp.2008.02.001

Nodine, C. F., & Kundel, H. L. (1987). Using eye movements to study visual search and to improve tumor detection. *RadioGraphics, 7*, 1241-1250. doi: 10.1148/radiographics.7.6.3423330

Olsen, K., Bahrami, B., Roepstorff, A., & Frith, C. (2010). Social interaction enhances the rate of individual visual perceptual learning. Poster presented at *MINDLab AU Symposium, Aarhus, Denmark, January 2011.* See http://bit.ly/Aal6px.

Palmer, E. M., Fencsik, D. E., Flusberg, S. J., Horowitz, T. S., & Wolfe, J. M. (2011). Signal detection evidence for limited capacity in visual search. *Attention, Perception, & Psychophysics, 73*, 2413-2424. doi: 10.3758/s13414-011-0199-2

Palmer, J., Verghese, P., & Pavel, M. (2000). The psychophysics of visual search. *Vision Research, 40*, 1227-1268. doi: 10.1016/s0042-6989(99)00244-8

Peirce, J. W. (2007). PsychoPy – Psychophysics software in Python. *Journal of Neuroscience Methods, 162,* 8-13.

Peirce, J. W. (2009). Generating stimuli for neuroscience using PsychoPy. *Frontiers Neuroinformatics, 2*, 10. doi: 10.3389/neuro.11.010.2008

Posner, M. I., Goldsmith, R., & Welton, K. E. (1967). Perceived Distance and The Classification of Distorted Patterns. *Journal of Experimental Psychology, 73*, 28-38.

Pratte, M. S., & Rouder, J. N. (2011). Hierarchical single- and dual-process models of recognition memory. *Journal of Mathematical Psychology*, *55*, 36-46. doi:10.1016/j.jmp.2010.08.007

Pratte, M. S., Rouder, J. N., & Morey, R. D. (2009). http://pcn.psychology.msstate.edu/

Pratte, M. S., & Rouder, J. N. (2010). Separating Mnemonic Process from Participant and Item Effects in the Assessment of ROC Asymmetries. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 224-232.

Pratte, M. S., & Rouder, J. N. (2012). Assessing the Dissociability of Recollection and Familiarity in Recognition Memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 1591-1607.

Pratte, M. S., Rouder, J. N., & Morey, R. D. (2010). Separating Mnemonic Process From Participant and Item Effects in the Assessment of ROC Asymmetries. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 36, 224-32. doi: 10.1037/a0017682.

Purvanova, R. K. (2014). Face-to-Face Versus Virtual Teams: What Have We Really Learned? *The Psychologist-Manager Journal, 17*, 2-29. doi: 10.1037/mgr0000009

Rains, S. A. (2005). Leveling the organizational playing field virtually: A meta-analysis of experimental research assessing the impact of group support system use on member influence behaviors. *Communication Research, 32*, 193-234. doi: 10.1177/0093650204273763

Reed, A.V. (1973). Speed-Accuracy Trade-Off in Recognition Memory. *Science,181,* 574-576. doi: 10.1126/science.1814099574.

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review, 12*, 573-604.

Schmidt, J., & Zelinsky, G. J. (2009). Search guidance is proportiona to the categorical specificity of a target cue. *Quarterly Journal of Experimental Research, 62*, 1904-1914. doi: 10.1080/17470210902853530

Scott, C. P. R., & Wildman, J. L. (2015). Culture, communication, and conflict: A review of the global virtual team literature. In *Leading global teams* (pp. 13-32). Springer, New York, NY.

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on Generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, *12*, 1123–1128. https://doi.org/10.1177/1745691617708630

Shachaf, P. (2008). Cultural diversity and information and communication technology impacts on global virtual teams: An exploratory study. *Information and Management, 45*, 131-142. doi: 10.1016/j.im.2007.12.003

Smith, B. N., Kerr, N. A., Markus, M. J. & Stasson, M. F. (2001). Individual differences in social loafing: need for cognition as a motivator in collective performance. *Group Dynamics: Theory, Research, and Practice, 5*, 150-158. doi: 10.1037//1089-2699.5.2.150

Smith, J. D., Redford, J. S., Gent, L. C., & Washburn, D. A. (2005). Visual Search and the Collapse of Categorization. *Journal of Experimental Psychology: General, 134*, 443-460. doi: 10.1037/0096-3445.4.443

Sorkin, R. D., & Dai, H. (1994). Signal Detection Analysis of the Ideal Group. *Organizational Behaviour and Human Decision Processes, 60*, 1-13. doi: 10.1006/obhd.1994.1072

Sorkin, R. D., Hays, C. J., & West, R. (2001). Signal-detection analysis of group decision making. *Psychological Review, 108*, 183-203. doi: 10.1037/0033-295X.108.1.183

Sorkin, R. D., West, R., & Robinson, D. E. (1998). Group performance depends on the majority rule. *Psychological Science, 9*, 456-463. doi: 10.1111/1467-9280.00085

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B: Statistical Methodology, 64*, 583-639.

Sproull, L., & Kiesler, S. (1991). New ways of working in the networked organization. *Administrative Science Quarterly, 37*, 491. doi: 10.2307/2393454

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behaviour Research Methods, Instruments, & Computers, 31*, 137-149. doi: 10.3758/bf03207704

Stasser, G., & Titus, W. (1985). Pooling of Unshared Information in Group Decision Making: Biased Information Sampling During Discussion. *Journal of Personality and Social Psychology, 48,* 1467-1478. doi: 0022-3514/85/$00.75

Steiner, I. D. (1972). *Group Process and Productivity.* New York: Academic.

Stroebe, W., Diehl, M., & Abakoumkin, G. (1996). Social compensation and the Köhler effect: Toward a theoretical explanation of motivation gains in group productivity. In E. Witte & J. Davis

(Eds.), *Understand- ing group behavior: Vol. 2. Small group processes and interpersonal relations* (pp. 37–65). Mahwah, NJ: Erlbaum.

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. *Choice Reviews Online, 42*, 42-1645. doi: 10.5860/choice.42-1645

Swets, J. A. (1986). Form of empirical ROCs in discrimination and diagnostic tasks: Implications for theory and measurement of performance. *Psychological Bulletin, 99,* 181-198. doi: 10.1037/0033-2909.99.2.181

Swets, J. A., Tanner, W. P., & Birdsall, T. G. (1961). Decision processes in perception. *Psychological Review, 68,* 301-340.

Tanner, W. P., & Swets, J. A. (1954). A decision-making theory of visual search. *Psychological Review, 61,* 401-409. doi: 10.1037/h0058700

Townsend, J. T., & Wenger, M. J. (2004). A Theory of Interactive Parallel Processing: New Capacity Measures and Predictions for a Response Time Inequality Series. *Psychological Review*, *111*, 1003–1035. https://doi.org/10.1037/0033-295X.111.4.1003

Triesman, A. (1985). Preattentive processing in vision. *Computer Vision, Graphics and Image Processing, 31*, 156-177. doi: 10.1016/0734-189x(85)90179-3

Triesman, A. M., & Gelade, G. (1980). A Feature-Integration Theory of Attention. *Cognitive Psychology, 12,* 97-136. doi: 10.1016/0010-0285(80)90005-5

Verghese, P. (2001). Visual Search and Attention: A Signal Detection Theory Approach. *Neuron, 31*, 523-535.

Vickery, T. J., King, L-W., & Jian, Y. (2005). Setting up the target template in visual search. *Journal of vision, 5*. doi: 10.1167/5.1.8.

Weber, B., & Hertel, G. (2007). Motivation gains of inferior group members: A meta-analytical review. *Journal of Personality and Social Psychology, 93,* 973-993

Wickelgren, W.A. (1977). Speed-Accuracy Tradeoff and Information Processing Dynamics. *Acta Psychologica, 41*, 67-85.

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science, 3,* 159-177. doi: 10.1080/14639220210123806

Wickens, C. D., Alexander, A. L., Ambinder, M. S., & Martens, M. (2004). The role of highlighting in visual search through maps, 373–388.

Wickham, H., & Chang, W. (2016). ggplot2: an implementation of the grammar of graphics, version 2.1.0. See https://cran. r-project. org/web/packages/ggplot2/index.html.

Williams, K. D., & Karau, S. J. (1991). Social loafing and social compensation: the effects of expectations of co-worker performance. *Journal of Personality and Social Psychology, 61*, 570-581. doi: 10.1037//0022-3514.61.4.570

Witte, E. H. (1989). Koehler rediscovered: the anti-Reingelmann effect. *European Journal of Social Psychology, 2*, 147-154. doi: 10.1002/ejsp.2420190206

Wittenbaum, G. M., & Stasser, G. (1996). Management information in small groups. In J. L. Nye & A. M. Brower (Eds.), *What's social about social cognition? Research on socially shared cognition in small groups* (pp.3-28). Thousand Oaks, CA, US: Sage Publications, Inc. doi: 10.4135/9781483327648.n1

Wolfe, J. M. (1994). Guided Search 2.0 A revised model of visual search. *Psychonomic Bulletin & Review, 1*, 202-238. doi: 10.3758/bf03200774

Wolfe, J. M., Horowitz, T. S., Van Wert, M. J., Kenner, N. M., Place, S. S., Kibbi, N. (2007). Low target prevalence is a stubborn source of errors in visual search tasks. *Journal of Experimental Psychology: General, 136*, 623-638. doi: 10.1037/0096-3445.136.4.623

Yonelinas, A. P., Dobbins, I., Szymanski, M. D., Dhaliwal, H. S., & King, L. (1996). Signal-Detection, Threshold, and Dual-Process Models of Recognition Memory: ROCs and Conscious Recollection. *Consciousness and Cognition*, *5*, 418-441. doi: 10.1006/ccog.1996.0026

Yu, R., & Wu, X. (2015). Working alone or in the presence of others: exploring social facilitation in baggage X-ray security screening tasks. *Ergonomics, 58*, 857-865. doi: 10.1080/00140139.2014.993429

Zajonc, R. B. (1965). Social Facilitation. *Science, 149*, 269-274. http://links.jstor.org/sici?sici=0036-8075%2819650716%293%3A149%3A3681%3C269%3ASF%3E2.0.CO%3B2-G