Attention, Perception, & Psychophysics (2020) 82:3387–3401
https://doi.org/10.3758/s13414-020-02087-y

Collaboration improves unspeeded search in the absence of precise target information

Alison Enright[1] · Nathan Leggett[1] · Jason S McCarley[2]

Published online: 8 July 2020
© The Psychonomic Society, Inc. 2020

Abstract
Two-person teams outperform individuals in search tasks, and even exceed expectations based on statistical limitations. Here, we aimed to replicate and extend this result. We used Bayesian hierarchical modelling of receiver operating characteristics to examine collaborative performance in a visual search task wherein top-down target information was constrained. Participants ($N = 16$ teams per experiment in Experiments 1 and 2; $N = 24$ teams in Experiment 3), working independently or collaboratively, performed a search task framed as a medical image reading task. Stimuli were polygons generated by randomly distorting a prototype shape. 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 (UW) model, one that assumed stochastically independent judgments and one that accounted for correlations in the team members' judgments. Collaborative search was better than that from single observers in all three experiments, and consistently trended higher than predictions of the correlated UW model. Results imply that collaborative search can be highly efficient even when target foreknowledge is limited.

Keywords Signal detection · Visual search · Bayesian modelling

Introduction

Signal detection theory (SDT; Green & Swets, 1966; Macmillan & Creelman, 2005) models decision-makers' ability to reach discrete judgments from uncertain data. A conventional signal detection task asks observers to distinguish two states of the world, one termed signal-plus-noise and the other noise-alone, on the basis of probabilistic evidence. In the standard SDT model, the psychological evidence distributions corresponding to signal-plus-noise and noise-alone states are normal with different means but the same standard deviation (Macmillan & Creelman, 2005). *Sensitivity* denotes the ability to discriminate signals from noise, and is typically 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 provides models for understanding collaborative or team decision making. Collaborative sensitivity in a signal detection task reflects the 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 averaging team members' individual judgments, weighting them according to each individual's mean sensitivity. Assuming an equal-variance Gaussian model and stochastic independence between team members' judgments, *d'* for the team is:

$$d'_{OW} = \sqrt{\sum d'^{2}_{i}}, \tag{1}$$

Public Significance Statement
Visual search is an important aspect of tasks such as medical image interpretation and transportation security screening, but can be inefficient, especially when top-down target knowledge is limited. Teams of searchers outperform an algorithm that averages the isolated searchers' judgments, indicating that collaboration improves efficiency even when top-down target knowledge is constrained.

✉ Alison Enright
alisonmargaretenright@gmail.com

[1] College of Education, Psychology and Social Work, Flinders University, GPO Box 2100, Adelaide, SA 5001, Australia

[2] School of Psychological Science, Oregon State University, Corvallis, OR, USA

where *i* indexes the team members. 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. Team *d'* under this model, again assuming an equal-variance Gaussian model stochastic independence between team members, is:

$$d^{\prime}_{UW} = \frac{\sum d'_i}{\sqrt{m}}, \tag{2}$$

where *m* is the number of team members. If team members are equally sensitive, the UW model is equivalent to the OW model. Otherwise, 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 six-member teams achieved higher sensitivity than individuals when recalling audiovisual information, but performed below the predictions of the UW/OW models. In contrast, Bahrami and colleagues (Bahrami et al., 2012; Bahrami et al., 2010) examined team performance in a two-interval, forced-choice visual search, and found sensitivity that matched the predictions of a UW model. This pattern held 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 that small teams (four or fewer members) in a multiple-cue judgment task performed with near-perfect efficiency, approaching predictions of the OW model. Efficiency decreased as team size increased, but this appeared to result from social loafing (Karau & Williams, 1993), rather than inefficient weighting strategies.

In the experiments conducted by Bahrami et al. (2010, 2012) and most of the experiments performed by Sorkin et al. (2001), importantly, stimuli were generated independently for each team member each trial. 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 radiologists both inspecting a chest radiograph for evidence of a lesion (Drew, Evans, Võ, Jacobson, & Wolfe, 2013; Kundel & Nodine, 1975). Although unique variance in their judgments might result from differences in the observers' sensory abilities, oculomotor scan patterns, or knowledge (Kundel & LaFollette, 1972; Nakashima et al., 2015; Nodine, Lauver, & Toto, 1996; Sowden, Davies, & Roling, 2000), properties of the image itself (Al Mousa et al., 2014) 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' judgments be known. Unfortunately, because the observers' mental decision variables are unobservable, the correlation between them might be difficult to estimate (Metz & Shen, 1992; but see Bollen & Barb, 1981; Enright, McCarley, & Leggett, 2019).

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 incorporate stimulus-driven dependencies in the decision-makers' judgments. Mock team judgments, in other words, reflect the collaborative sensitivity that can be expected given the stimulus-driven correlations between the team members' judgments.

Taking this approach, 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) or alone (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 showed greater sensitivity than did mock teams, indicating 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 to one half of the search field and second team member attending to the other half.

More recently, Enright et al. (2019) examined collaborative performance in a mock baggage screening task. Participants viewed simulated baggage x-rays and judged whether a knife, drawn from a set of five potential items, was present in each trial. In the individual search condition of Experiment 1, the participants performed the task in isolated rooms. In the collaborative condition, they sat side-by-side and were allowed to discuss each 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, incorporated stimulus-driven dependencies by averaging individual team members' ratings of yoked stimuli to create team judgments. The other, denoted the $UW_{\rho\,=\,0}$ model, assumed independent judgments, and predicted team sensitivity using Equation 2. 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'* that does not require the assumption of equal-variance evidence distributions.

Echoing the findings of Malcolmson et al. (2007), teams achieved higher sensitivity than single observers, producing mean d'_e values that fell 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 task while communicating via speakerphone, and when viewing times were restricted. Two interpretations of the results were considered. The first was that the collaborative gains reflected social compensation, with observers putting forth more effort when they worked collaboratively than when they worked alone. The second was that teams might have interacted in a way that improved their information sampling as they viewed the images. Most obvious was the possibility that searchers adopted a division-of-labor strategy (Malcolmson et al., 2007), either by making individual team members responsible for inspecting particular regions of the display, or alternatively, by making individual team members responsible for detecting particular targets (for evidence of similar strategies in speeded collaborative search, see Brennan, Chen, Dickinson, Neider, & Zelinsky, 2008; Niehorster, Cornelissen, Holmqvist, & Hooge, 2019 and Wahn, Czeszumski, Labusch, Kingstone, & König, 2020; though see Yamani, Neider, Kramer, & McCarley, 2017, for evidence of increased overlap in oculomotor scanning in a team search, and McCarley, Leggett, & Enright, 2020, for behavioral evidence of inefficient team performance in a visual monitoring task).

The present experiments aimed to test the division-of-labor hypothesis by using the procedure of Enright et al. (2019) with stimuli designed to restrict potential division-of-labor strategies. In our previous experiments, pictures of the five potential targets were provided with the task instructions, and the participants performed a set of practice trials before beginning the experimental trials. Thus, the space of potential target items was small and familiar. This would have made it easy for collaborators to divide the display space while searching for any of the potential targets, or alternatively, to divide the set of targets while searching across the display.

However, many naturalistic tasks require observers to search for targets whose appearance is uncertain. In the example of radiologists given above, for instance, the chest radiograph might contain a lesion whose shape and size are not known *a priori*. Because top-down control guides attention towards items that match the target's features, search is generally less efficient when the target is uncertain (Chen & Zelinsky, 2006; Wolfe, 1994), and is most efficient if observers are provided a detailed and accurate template of the target (Hout & Goldinger, 2014; Malcolm & Henderson, 2009; Schmidt & Zelinsky, 2009; Vickery, King, & Jiang, 2005).

Smith and colleagues (Smith, Redford, Gent, & Washburn, 2005) found especially poor performance in a novel form of visual search characterized by weak top-down control. Stimuli were polygons created by randomly distorting prototype shapes (Posner, Goldsmith, & Welton, 1967). Participants searched 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 guessing 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 modified the task and stimuli of Smith et al. (2005) to introduce further target uncertainty, limiting the prospects of a division-of-labor strategy. Here, all the objects within a given stimulus image were distortions of a common prototype. The distractors, though, were distorted only modestly. The target, when it was present, was distorted to a larger degree. Thus, the target could be distinguished only by comparison to the surrounding distractors. This design limited the possibility that collaborators could divide responsibility either by searching different regions of the display, or by searching for different items within the set of potential targets. 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. Comparing search stimuli in this way inherently limited the ability of team members to restrict search to predefined search areas.

Building on Enright et al. (2019), we use Bayesian hierarchical analysis of the ROC (Morey, Pratte, & Rouder, 2008; Pratte, Rouder, & Morey, 2009; Pratte & Rouder, 2012) to examine collaborative visual search using distorted-polygon stimuli. The search was framed as a medical image reading task. Participants were told their task was to inspect cell samples for an abnormal cell, working either 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 $\mathrm{UW}_{\rho = 0}$ 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 and analyses were preregistered (Enright, McCarley, & Leggett, 2018a, 13 June) (https://osf.io/u43yg/?view_only=78d0a609a9b64dcb8d1b9c70094c6dc3).

Method

Participants

Sixteen pairs of undergraduate students (22 female, M_{age} = 22.5 years, *SD* = 3.28) were recruited via Finders University College of Education, Psychology, and Social Work's first-year research participant pool. Our sample size was chosen to match that of the experiments of our previous study (Enright et al., 2019), which found consistent results across a series of five replications. All participants demonstrated normal or corrected-to-normal visual acuity and color vision and were paid $20AU in exchange for participation. This experiment was approved by the Social and Behavioural Research Ethics Committee (SBREC) at Flinders University.

Apparatus and stimuli

Participants completed the visual search task on a 370 mm × 300 mm Samsung monitor (model S24D590PL), with a resolution of 1,920 × 1,080 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 570 mm, though viewing distance was not constrained.

Distorted polygon stimuli were generated using the R programming language (R Core Team, 2018). Each stimulus image comprised a set of three to four polygons created by distorting a common prototype. A prototype was generated by randomly selecting a sequence of five points within a 30 × 30 (21° × 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., 1967). All objects were rendered as colored regions drawn at 50% opacity. The color of each item was selected randomly with replacement from the default color palette in R. Each item was positioned randomly, under the constraint that the object did not extend beyond the bounds of an imaginary 6° × 6° box concentric with the center of the display.

Stimulus images were generated in yoked target-absent/target-present pairs. The target-absent image within a pair contained only distractors. The yoked target-present image was identical, except that one distractor was replaced with a target, centered at the same position and drawn in the same color. Figure 1 presents a pair of yoked images. A total of 400 pairs of images were generated, and were sorted randomly into two sets, A and B, of 200 pairs each.

Procedure

Procedure was similar to that of Enright et al. (2019). Participants completed the visual search task in the same testing room in all conditions. In the single-observer condition, participants worked independently, sitting at workstations at a 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, the participants sat side-by-side at one workstation.

Instructions were presented onscreen at the start of the experimental session, and framed the search as a mock cell pathology screening task, explaining that the participants' 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 1,000-ms fixation phase. Participants then viewed the stimulus image and confidence rating scale, presented below the stimulus image, freely, until a response was made. Participants executed responses by mouse clicking one of six response options that included *Definitely yes, Probably yes, Guess yes, Guess no, Probably no,* and *Definitely no*. After a response was executed, participants were informed about the accuracy of their judgments via feedback messages presented onscreen. Specifically, hits produced a message reading, "You found a highly abnormal cell!", and correct rejections produced a message reading, "Good judgment." Misses were followed by, "You missed a highly abnormal cell!" and false alarms were followed by, "False alarm." *Definitely yes, probably yes,* and *guess yes* were treated as correct responses for target present trials and, similarly, *definitely no, probably no,* and *guess no* were treated as correct responses for target absent trials.

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 the single observer conditions and set B for the team search conditions. The remaining teams used set B for the single observer condition and set A for team search. Trial order was randomized within blocks and yoked across participants in the single observer condition.

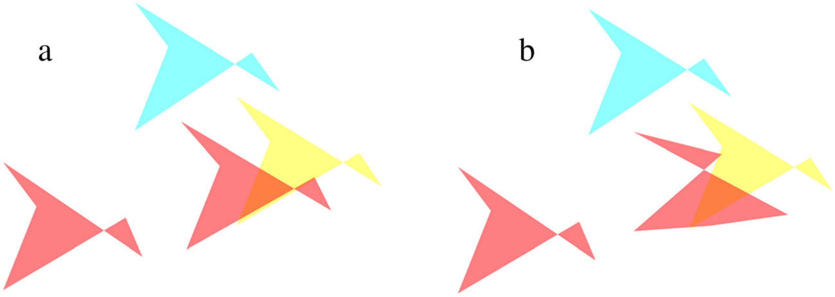

Fig. 1 An example of a yoked stimulus pair. **Left panel:** Target-absent image. **Right panel:** Target-present image

Analyses

Data analysis was performed using the HBMEM (Pratte, 2018) package in R statistical software. The HBMEM package was created to fit signal detection models to confidence-rating data (Morey et al., 2008; Pratte et al., 2009; Pratte & Rouder, 2012) and accommodates hierarchical versions of both the equal and unequal variance Gaussian forms of detection models. The model is fit with a Bayesian Markov chain Monte Carlo (MCMC) sampling procedure, using vague priors on model parameters. The model-fitting procedure was run for 10,000 burn-in iterations and 50,000 iterations for analysis.

The model assumes a noise distribution with a mean of 0 and standard deviation of 1, and estimates parameters of the signal-plus-noise distribution. Parameters are estimated hierarchically, with the parameters at the unit (individual or team) level sampled from group-level distributions. We fit three versions of the model to judge which model produced the best fit. First, the equal variance model (EV) assumed the signal-plus-noise and noise-alone distributions had the same variance. Second, the unequal variance fixed S^2 (UV, fixed S^2) allowed that the variance in the signal-plus-noise distribution might differ from that in the noise-alone distribution, but assumed that the signal-plus-noise variance was fixed across observers. Third, the unequal variance free S^2 (UV, free S^2) allowed the variance of the signal-plus-noise distribution to vary across observers (Pratte & Rouder, 2012). The quality of model fits was determined using the deviance information criterion (DIC). DIC provides a measure of model fit, adjusted for the number of model parameters (Spiegelhalter, Best, Carlin, & van der Linde, 2002). Lower DIC values indicate better model performance.

A typical paired-samples comparison of search condition (team, individual searcher) was not possible because two single searchers comprised each team. As such, search condition was treated as a between-subject variable with 32 participants in the single observer condition and 16 participants in the team search condition. Before model fitting, data were checked to confirm that all individuals and teams met a preregistered inclusion criterion of 60% accuracy or better.

Predicted d'_e scores from the UW$_{\rho=0}$ model were generated using the hierarchical group mean parameter estimates of μ_s and σ_s for the single-observer condition at each iteration of the MCMC process. Because the modeling provided only one group-level estimate of each parameter in the single observer condition – 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 the OW model. The mock UW model predictions were generated by first averaging the two searchers' responses on yoked trials of the single observer condition, truncating the result to place it on a six-point scale, and finally submitting the ratings to the HBMEM model.

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

Results

All participants and teams in Experiment 1 met the minimum 60% accuracy level for inclusion. The DIC values for Experiments 1, 2, and 3 are presented in Table 1, and show that the UV, S^2 free produced the best fit, adjusted for the number of model parameters, in all three experiments. Only the results of that model are therefore reported below. Trace plots of the MCMC chains for single observers, teams, and

Table 1 Deviance information criterion (DIC) values for the equal variance (EV) SD, unequal variance (UV)SD S^2 fixed and UVSD S^2 free for Experiments 1 and 2

Experiment	DIC values		
	EV	UV, S^2 fixed	UV, S^2 free
1	26,538.85	26,543.86	26,467.20
2	26,551.16	26,505.11	26,484.08
3	42,239.81	42,222.86	42,167.94

mock UW model predictions for Experiment 1 are shown in Fig. 2. Chains vary narrowly, suggesting that they have settled within ranges of values representative of the true posterior distribution (Kruschke, 2015).

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 Fig. 3. The z-slopes for the signal-plus-noise and noise-alone 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 4 shows the posterior distribution of d'_e scores for single observers, teams, the $UW_{\rho=0}$ model, and the mock UW model, based on estimates of the group-level parameters. Figure 5 shows the distributions and 95% BCIs of the difference scores between observed and predicted team performance levels. As expected, teams, M = 2.21, BCI [1.98, 2.44], outperformed single observers, M = 1.71, BCI [1.57, 1.85], M_{diff} = 0.5, BCI [0.24, 0.77]. Team performance was again numerically worse than the $UW_{\rho=0}$ model's predictions, though the difference did not reach the 95% credible level, M = 2.41, BCI [1.66, 2.08], M_{diff} = -0.21, BCI [-0.51, 0.09]. More importantly, teams outperformed mock UW teams, M = 1.86, BCI [1.66, 2.08], M_{diff} = 0.34, BCI [0.03, 0.65].

Discussion

Empirical teams achieved higher sensitivity than mock teams based on individual searchers' averaged response ratings, outperforming the predictions of a UW rule applied to participants' stochastically dependent judgments. This pattern mimics the results of Enright et al. (2019) and extends the findings to a task characterized by target uncertainty. Experiment 2 aimed to replicate and extend these findings, repeating the procedure of Experiment 1 with 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 through voice chat software. Our earlier study (Enright et al., 2019) found that teams outperformed mock teams in a visual search task even when team members worked in separate rooms, communicating by speech. The current experiment tested whether this would hold true even when target certainty was highly limited. Experimental methods were preregistered (Enright,

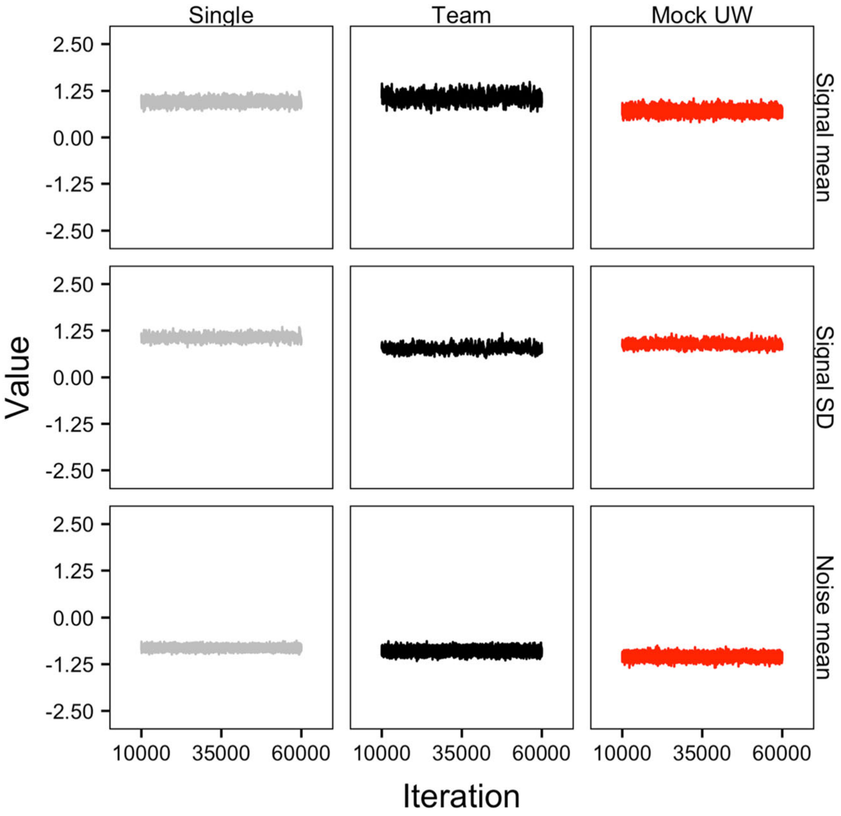

Fig. 2 Markov Monte Carlo (MCMC) chains in Experiment 1. Rows show estimated parameters and columns the search condition

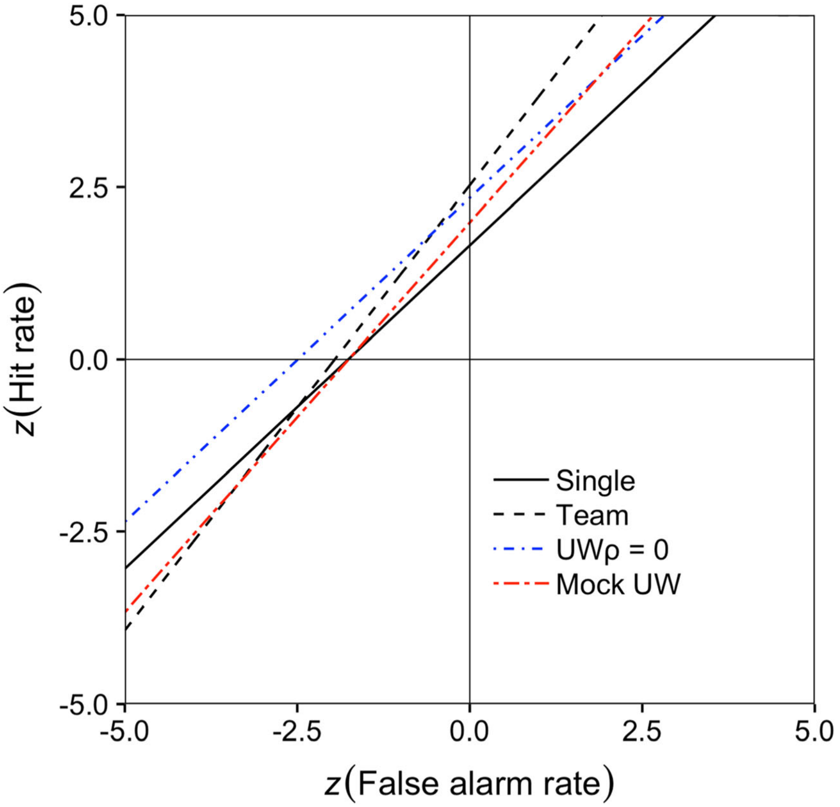

Fig. 3 zReceiver operating characteristics (ROCs) for single observers (solid black line), teams (black dashed line), uniform judgment-weighting $(UW)_{\rho=0}$ model-predicted (blue dashed line), and mock UW model-predicted (red dashed line) performance in Experiment 1

McCarley, & Leggett, 2018b, 13 June) (https://osf.io/wcp69/?view_only=2069ee7afc434eef901820006a5f6665).

Method

Thirty-two participants, making 16 pairs, of undergraduate students (23 female, M_{age} = 22.25 years, *SD* = 8.80) were recruited, screened, and remunerated in the same manner as in Experiment 1. This experiment was approved by the Social and Behavioural Research Ethics Committee (SBREC) at Flinders University.

All stimuli and procedures were exactly the same as Experiment 1, except that participants performed both the individual and collaborative conditions in separate testing rooms. When performing the collaborative condition, team members communicated via Skype (www.skype.com) using only the phone function (i.e., no video). The Skype software operated on the same computer as the visual search task but was not visible during the search task.

Results

All individual participants and teams in Experiment 2 met the minimum 60% accuracy inclusion criteria. Figure 6 shows the trace plots of the MCMC chains for single observers, teams, and mock UW model predictions for Experiment 2. By inspection, chains vary narrowly, suggesting they have settled within regions representative of the true posterior.

Figure 7 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 8 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 [1.82, 2.56], M_{diff} = 0.39, BCI [0.15 0.64]. Team sensitivity trended higher than mock UW teams, *M* = 2.11, BCI [1.91, 2.32]), though the credible interval on the difference did not exclude 0, M_{diff} = 0.23, BCI [-0.06, 0.52] (see Fig. 9). Team performance fell below the level predicted by the $UW_{\rho=0}$ model, *M* = 2.76, BCI [2.57, 2.95], M_{diff} = -0.41, BCI [-0.69, -0.13].

Discussion

As predicted, empirical teams outperformed single observers and collaborative sensitivity fell midway between the predictions of the correlated and uncorrelated UW model

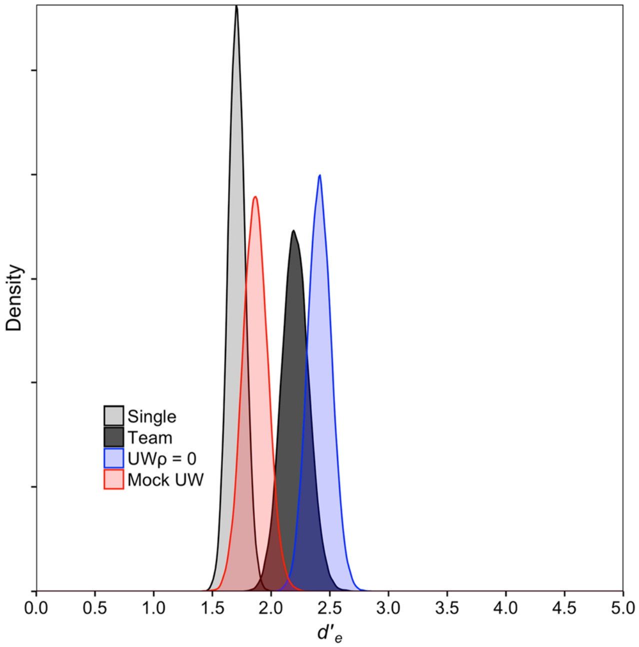

Fig. 4 The posterior distributions of d'_e for single observers (light gray), teams (dark gray), mock uniform judgment-weighting (UW) teams (red), and the UW$_{\rho=0}$ model (blue) in Experiment 1

predictions, though it was not credibly different from mock-UW model predictions. These effects roughly replicate those obtained in Experiment 1 and extend them to a context in which teams were not co-located.

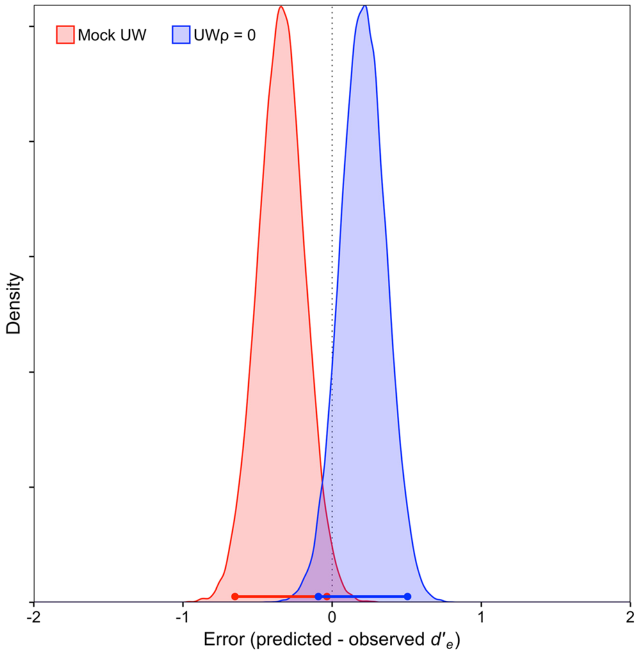

Fig. 5 The 95% Bayesian credible intervals (BCIs) of the difference scores between observed team performance and the Mock uniform judgment-weighting (UW) model (red) and the UW$_{\rho=0}$ model predictions (blue) in Experiment 1

Experiment 3 asked whether better-than-expected team performance observed in Experiments 1 and 2 might reflect a speed-accuracy trade-off.

Experiment 3

The procedure of Experiment 3 was identical to that of Experiment 1 except that the stimuli were presented onscreen for a brief duration. In both the experiments above, teams (M_{EXPT1} = 2.35s, M_{EXPT2} = 8.82s) took longer to respond than did individuals (M_{EXPT1} = 1.44s, M_{EXPT2} = 2.01s). These differences likely resulted in part from the time needed to discuss and reach a joint decision but might also reflect a speed-accuracy trade-off (Reed, 1973; Wickelgren, 1977) whereby teams scanned the stimulus images longer than individuals did. Although our earlier experiments (Enright et al., 2019) found no evidence that increased scanning times contributed to collaborative sensitivity, the possibility of a speed-accuracy trade-off remains open in the current task. Experiment 3 attempted to rule this possibility out. Methods were preregistered (Enright et al., 2019, 4 June) (https://osf.io/wcp69/?view_only=2069ee7afc434eef901820006a5f6665).

Method

Twenty-four teams, 48 participants (35 female, M_{age} = 22.65 years, *SD* = 4.27), experienced identical recruitment, screening and renumeration as Experiment 1. Experiment 3 was approved by the Flinders University Social and Behavioural Research Ethics Committee (SBREC).

The stimuli and procedure of Experiment 3 were identical to those of Experiment 1 except that the stimuli were presented onscreen for 1,000 ms, after which participants had an unlimited time to execute a response. Assuming that the mean response time for single observers in Experiments 1 and 2 includes time to both scan images and execute a response, 1,000 ms provides a conservative estimate of time spent scanning images only.

Results

All single observers and teams met the minimum 60% accuracy inclusion criteria. Figure 10 shows trace plots of the MCMC chains for single observers, teams, and mock UW model predictions. By inspection, chains vary narrowly, suggesting they have settled within regions representative of the true posterior.

The *z*ROCs for the single observers, teams, UW$_{\rho=0}$ model and mock UW team model predictions, based on estimates of the group-level parameters, are shown in Fig. 11. The signal-plus-noise and noise-alone distributions' z-slopes were less than 1.0 (*M* = 0.82 for single observers and *M* = 0.79 for

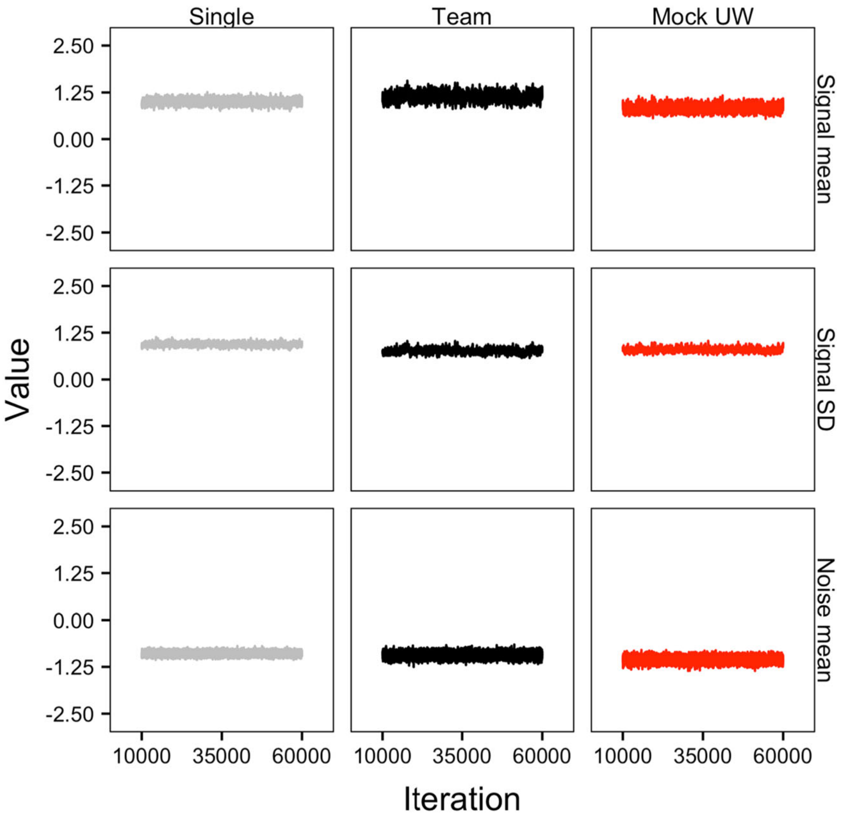

Fig. 6 Markov Monte Carlo (MCMC) chains in Experiment 2. The rows show estimated parameters and the columns show search condition

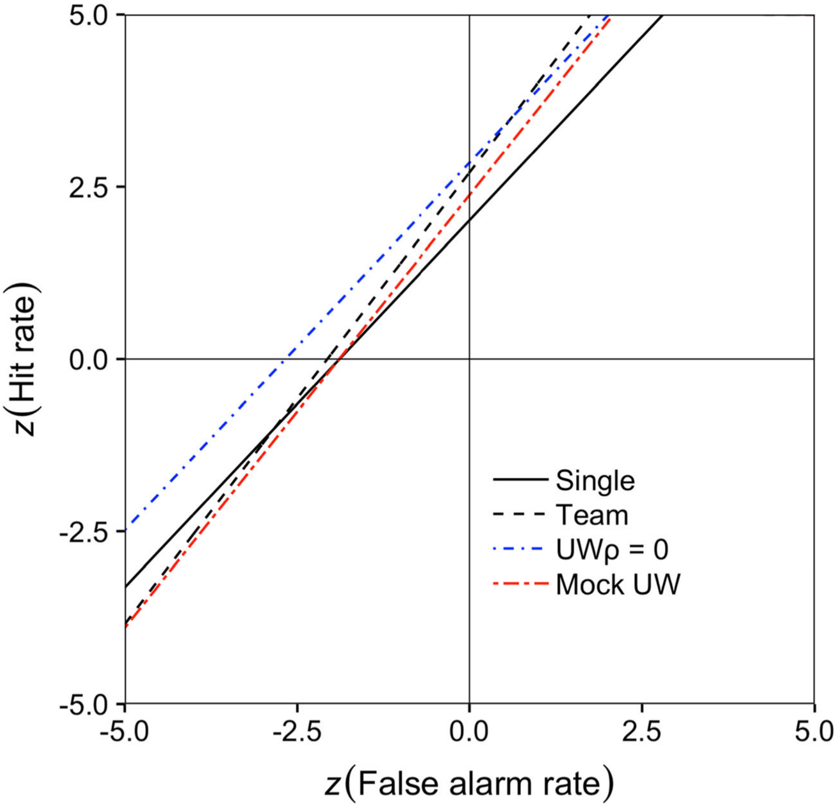

Fig. 7 *z* Receiver operating characteristics (ROCs) for single observers (solid black line), teams (black dashed line), uniform judgment-weighting $(\mathrm{UW})_{\rho=0}$ model-predicted (blue dashed line), and mock UW model-predicted (red dashed line) performance in Experiment 2

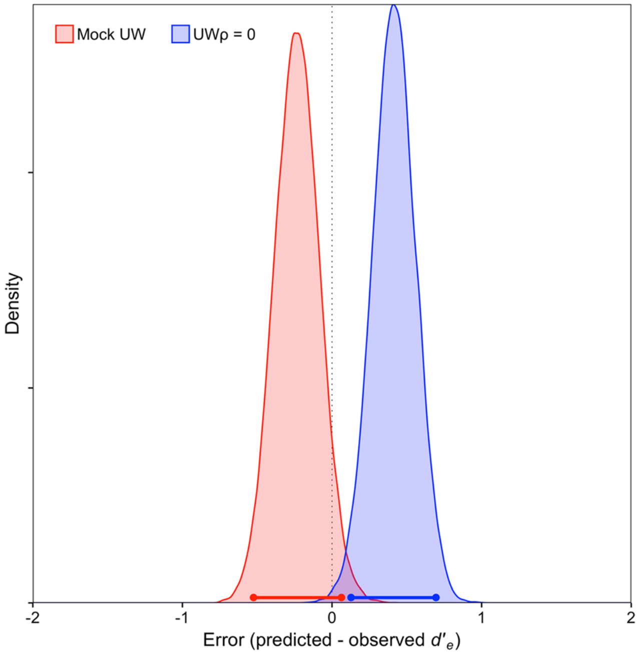

Fig. 9 The 95% Bayesian credible intervals (BCIs) for the differences scores between observed team performance and Mock uniform judgment-weighting (UW) model (red) and UW$_{\rho=0}$ model (blue) predictions in Experiment 2

teams) indicating that the signal-plus-noise distribution had a larger variance than the noise-alone distribution.

The posterior distribution of d'_e scores for single observers, teams, UW$_{\rho=0}$ model-predicted and mock UW model-predicted performance are shown in Fig. 12. Teams, M = 2.08, BCI [1.93, 2.24], outperformed single observers, M = 1.71, BCI [1.61, 1.81], M_{diff} = 0.38, BCI [0.19, 0.56]. As in Experiments 1 and 2, team sensitivity trended higher than mock UW teams, M = 1.89, BCI [1.74, 2.05]); however, the credible interval on the difference did not exclude 0, M_{diff} = 0.19, BCI [-0.03, 0.41] (see Fig. 13). Team performance was again below that predicted by the UW$_{\rho=0}$ model, M = 2.41, BCI [2.27, 2.56], M_{diff} = -0.33, BCI [-0.54, -0.12].

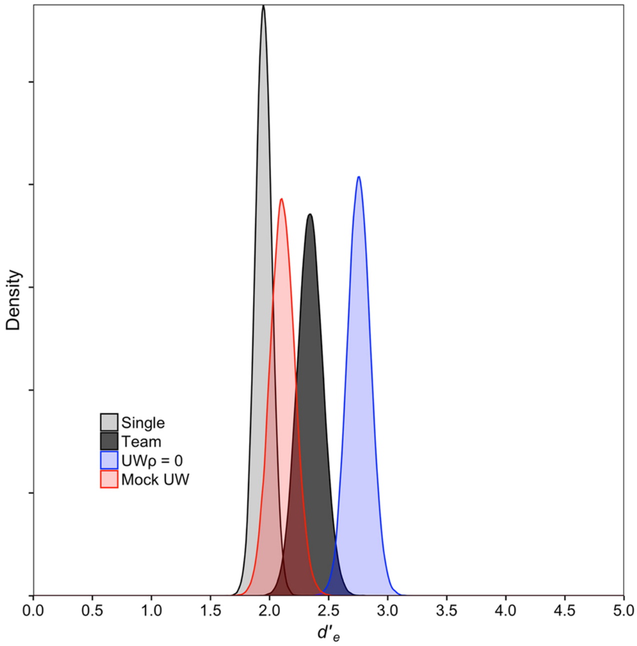

Fig. 8 The posterior distributions of d'_e for single observers, teams, mock uniform judgment-weighting (UW) teams, and the UW$_{\rho=0}$ model in Experiment 2

Meta-analysis

Although the results of specific comparisons between conditions differed, Experiments 1, 2, and 3 produced generally similar effects. To provide a better estimate of the effect sizes, a final analysis combined the data of the three experiments. The combined data were fit using the same model as in the three experiments above.

Figure 14 presents the estimated posterior distributions of d'_e. Teams, M = 2.19, BCI [2.08, 2.30] outperformed individual searchers, M = 1.77, BCI [1.70, 1.84], M_{diff} = 0.42, BCI [0.28, 0.55]. Team sensitivity fell midway between mock team sensitivity, M = 1.94, BCI [1.84, 2.05], M_{diff} = 0.25, BCI [0.09, 0.40], and the predicted sensitivity of the UW$_{\rho=0}$ model, M = 2.51, BCI [2.41, 2.61], M_{diff} = - 0.32, BCI [-.047, -0.17], and was credibly different from both (see Fig. 15).[1]

General discussion

Three experiments examined collaborative visual search in a task characterized by high target uncertainty. Observed team performance was compared to the predictions of two versions of a uniform weighting model – one that incorporated the stochastic dependency between team members' judgments and one that assumed stochastic independence. 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 Experiment 3, teams performed the task in the same testing room but experienced only a brief stimuli exposure duration. As expected, teams in all three experiments performed better than single observers. More importantly, empirical team performance fell short of the predictions of the UW$_{\rho=0}$ model but exceeded

[1] A reviewer noted that even without top-down knowledge of target shape, stimuli with three nontargets and one target might allow a spatial division of labor strategy, since two distractors and the target could sometimes fall within a region of the display monitored by just one participant. To test whether this hypothesis might explain the collaborative benefits observed across our experiments, we reran our meta-analysis using only the data for trials on which the stimulus contained three items (three distractors or two distractors and a target). The results of this analysis matched the pattern of results obtained with the full data set, suggesting that a spatial division-of-labor strategy with set size four did not substantially inflate collaborative gains. We thank the reviewer for raising this issue.

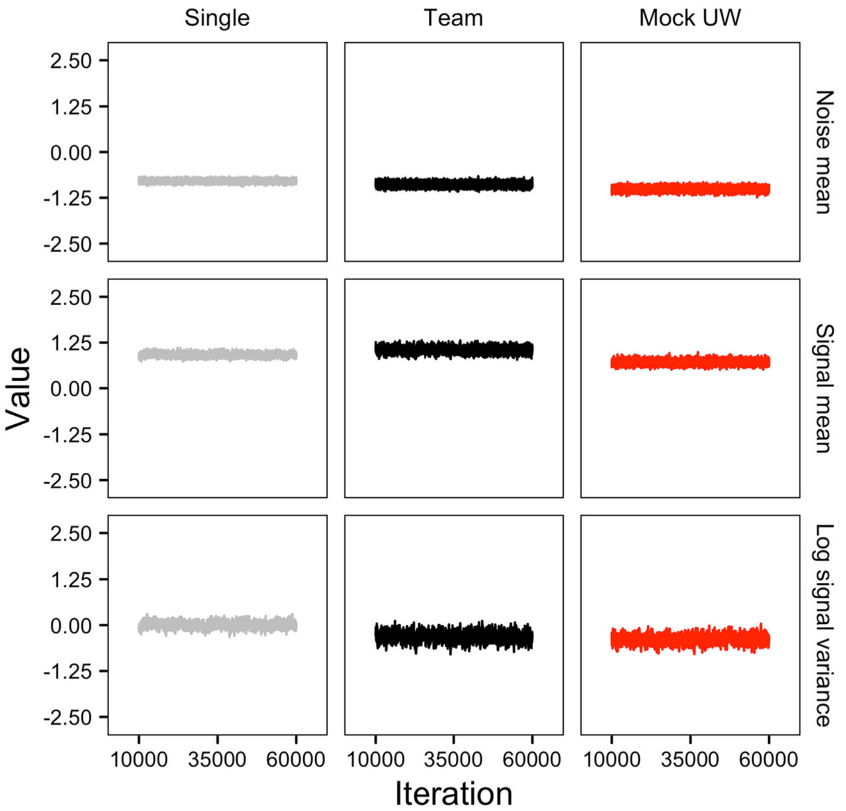

Fig. 10 Experiment 3 Markov Monte Carlo (MCMC) chains. The rows show estimated parameters and the columns show search condition

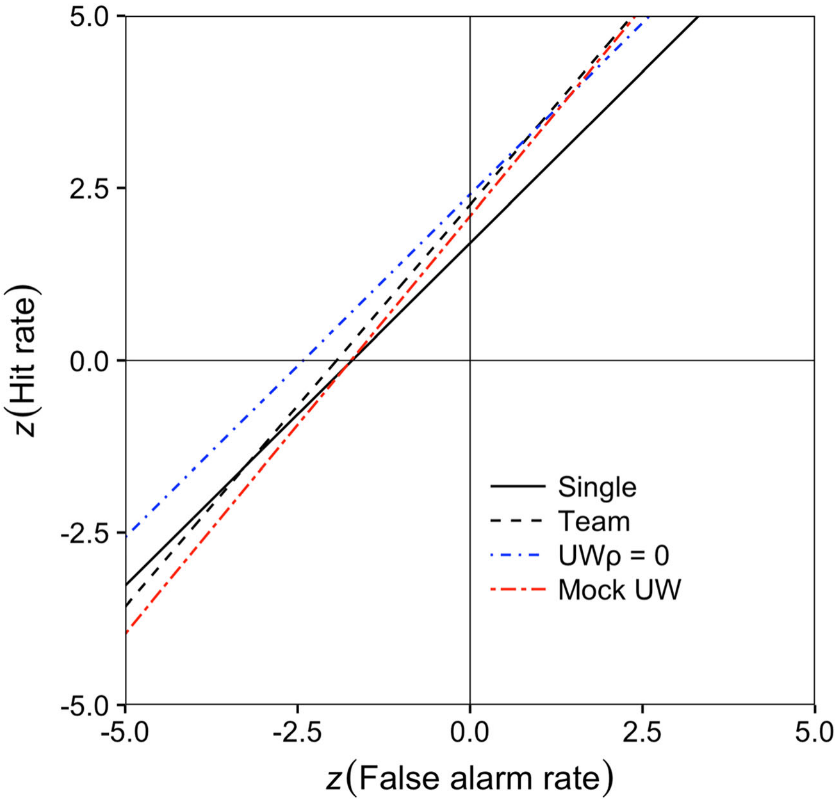

Fig. 11 z Receiver operating characteristics (ROCs) for single observers (solid black line), teams (black dashed line), uniform judgment-weighting (UW) $_{\rho=0}$ model-predicted (blue dashed line), and mock UW model-predicted (red dashed line) performance in Experiment 3.

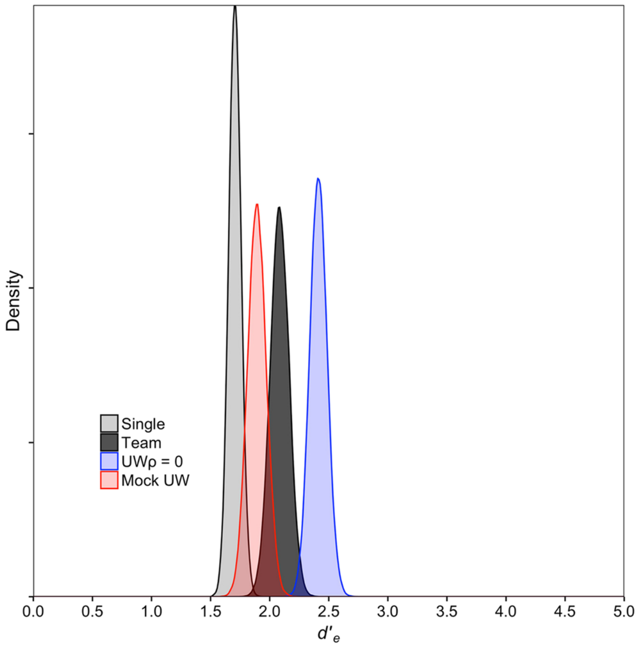

Fig. 12 The posterior distributions of d'_e for single observers, teams, mock uniform judgment-weighting (UW) teams, and the $UW_{\rho=0}$ model in Experiment 3

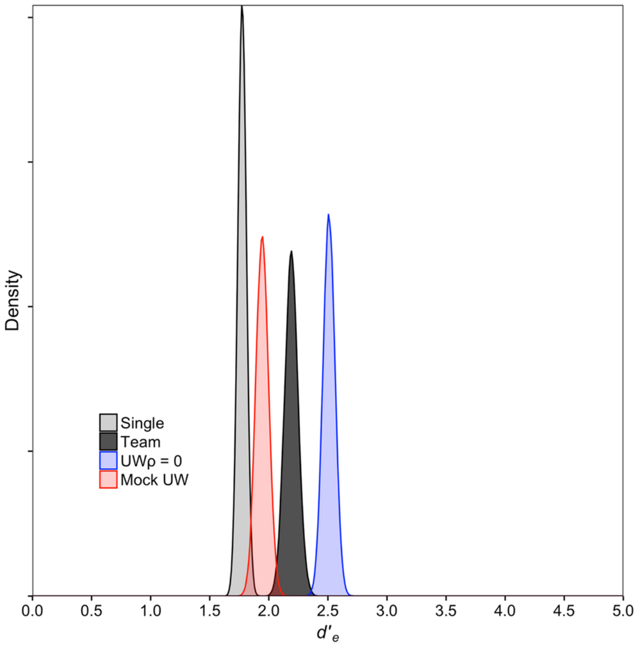

Fig. 14 The posterior distributions of d'_e for single observers, teams, mock uniform judgment-weighting (UW) teams, and the $UW_{\rho=0}$ model in the meta-analysis

mock team performance (credibly in Experiment 1 and in the meta-analysis). In other words, empirical teams achieved higher sensitivity than expected from a UW strategy, given the correlation between team members' judgments.

Results replicate Enright et al. (2019) findings and extend them to a search task that limits top-down knowledge of the target. Here, the items within a given stimulus image were all probabilistic distortions of a common prototype shape, and the target differed from the distractors only in the degree of distortion. Participants were aware that targets would be distorted versions of the distractor items, but were unaware of what specific shape the target might present. This stimulus and task

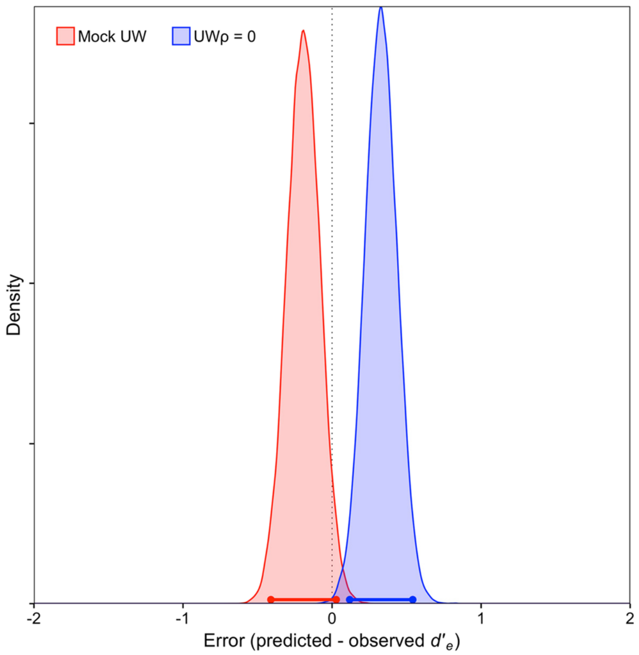

Fig. 13 The 95% Bayesian credible intervals (BCIs) of the difference scores between observed team performance and the Mock uniform judgment-weighting (UW) model (red) and the $UW_{\rho=0}$ model predictions (blue) in Experiment 3

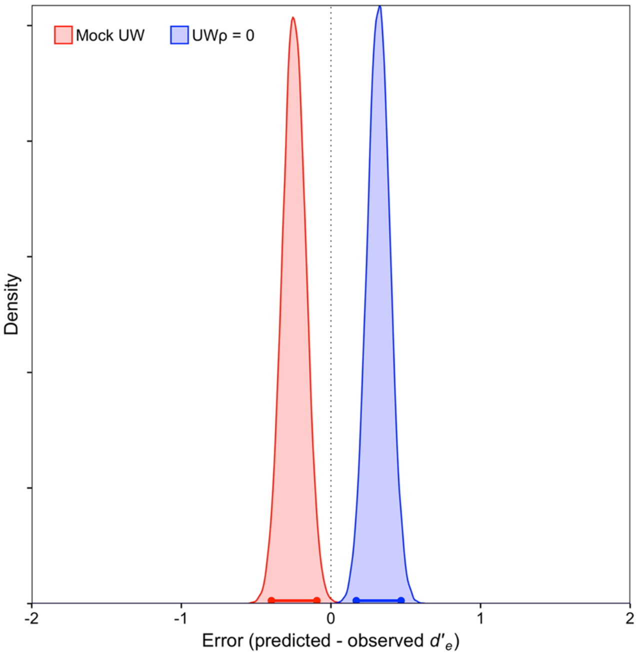

Fig. 15 The 95% Bayesian credible intervals (BCIs) of the difference scores between observed team performance and the Mock uniform judgment-weighting (UW) model (red) and the $UW_{\rho=0}$ model predictions (blue) in the meta-analysis

aimed to make division-of-labor strategies, of the sort that have been proposed to explain team performance gain in earlier research impracticable. Past work has suggested that paired searchers might optimize team performance by dividing the display space between them, with each teammate searching one region of the display (Malcolmson et al., 2007; cf. Brennan et al., 2008). Alternatively, given a limited set of potential target stimuli, they might optimize performance by dividing the target space between them, each teammate searching for a specific subset of the target items (Enright et al., 2019).

In the current task, though, neither of those strategies is likely to have been successful. Because the target shape of each trial was unknowable *a priori*, participants could not adopt an attentional set for a particular target. And because the target could be identified only by comparison to the distractors within the display, searchers would have been required to attend to at least three items each trial to judge whether or not a target was present. A strategy of restricting attention to a limited region of the display is therefore unlikely to have been helpful. The present data thus suggest that searchers may use collaborative strategies beyond a division of the display or target space.

What strategy might participants have used instead? Because viewing times were unlimited (in Experiments 1 and 2), a potential concern is that the teams might have simply spent more time searching than individuals, producing a speed-accuracy trade-off that inflated team sensitivity. Experiment 3 ruled out this possibility by fixing stimuli exposure duration to a brief onscreen presentation and found results very similar to the patterns observed in Experiments 1 and 2. These findings fit nicely with our previous work (Enright et al., 2019) that showed that even when stimulus exposure duration was limited, and matched across individual and team search conditions, teams continued to outperform the sensitivity levels predicted by mock teams. Thus, exposure duration doesn't seem necessary for teams to outperform mock teams.

A different possibility is that searchers might have put forth more effort when working collaboratively than when working alone, increasing their individual sensitivity before integrating their judgments to reach a joint decision (Kerr & Tindale, 2004). Alternatively, team members might have acted as mutually facilitatory channels (Eidels, Houpt, Altieri, Pei, & Townsend, 2011), exchanging information in a way that allowed each of them to accrue evidence more effectively than when working alone. Future research, using eye tracking or analyzing verbal protocols, might help distinguish these possibilities.

Regardless of their explanation, the current data indicate that teams outperform mock teams even when target properties are not well specified *a priori*. This implies that collaborative search can improve performance in naturalistic tasks such as the baggage x-ray screening or the detection of anomalies in visual images (Kurtz & Gentner, 2013), tasks in which target foreknowledge is often imperfect at best.

Generalizing the current findings is limited to the characteristics of the sample population and the task (Simons, Shoda, & Lindsay, 2017). The experimental task was carefully controlled but artificial, using highly abstract, computer-generated stimuli, and participants were students performing the task for the first time. Additional research will be necessary to generalize the current results to expert observers (e.g., radiologists) performing a more naturalistic task. Participants also performed the task in a quiet room with no distractions and no time stress. Results may differ in more naturalistic contexts characterized by distractions and noise.

Author Note This research is supported by ARC Discovery DP150100661 (the Australian Research Council) awarded to Jason S. McCarley.

References

Al Mousa, D. S., Brennan, P. C., Ryan, E. A., Lee, W. B., Tan, J., & Mello-Thoms, C. (2014). How mammographic breast density affects radiologists' visual search patterns. *Academic Radiology*, *21*, 1386–1393.

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

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: https://doi.org/10.1098/rstb.2011.0420

Bollen, K. A., & Barb, K. H. (1981). Pearson's R and Coarsely Categorized Measures. *American Sociological Review*, *46*, 232–239. https://doi.org/10.2307/2094981

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

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

Drew, T., Evans, K., Võ, M. L.-H., Jacobson, F. L., & Wolfe, J. M. (2013). Informatics in radiology: What can you see in a single glance and how might this guide visual search in medical images? *RadioGraphics*, *33*, 263–274. https://doi.org/10.1148/rg.331125023

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*(6), 768-773.

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

Enright, A., & McCarley, J. S. (2019). Collaborative Search in a Mock Baggage Screening Task. J*ournal of Experimental Psychology: Applied.* https://doi.org/10.1037/xap0000216

Enright, A., McCarley, J. S., & Leggett, N. (2018a). Benchmarking Collaborative Search using a Bayesian Hierarchical Regression Analysis. Retrieved from osf.io/u43yg

Enright, A., McCarley, J. S., & Leggett, N. (2018b). Benchmarking Collaborative Search using a Bayesian Hierarchical Regression Analysis. Retrieved from osf.io/wcp69

Enright, A., McCarley, J. S., & Leggett, N. (2019) Benchmarking Collaborative Search using a Bayesian Hierarchical Regression Analysis. Retrieved from osf.io/qhpx6

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

Hinsz, V. B. (1990). Cognitive and consensus processes in group recognition memory performance. *Journal of Personality and Social psychology, 59*(4), 705.

Hout, M. C., & Goldinger, S. (2014). Target templates: the precision of mental representations affects attentional guidance and decision-making in visual search. *Attention, Perception, and Psychophysics, 77*(1), 128-149. https://doi.org/10.3758/s13414-014-0764-6

Karau, S. J., & Williams, K. (1993). Social Loafing: A Meta-Analytic Review and Theoretical Integration. *Journal of Personality and Social Psychology, 65*(4), 681-706. https://doi.org/10.4135/9781412984997.n7

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

Kruschke, J. K. (2015). *Doing Bayesian data analysis* (2nd). Academic Press/Elsevier.

Kundel, H. L., & LaFollette, P. S. (1972). Visual search patterns and experience with radiological images. *Radiology, 103*, 523-528.

Kundel, H. L., & Nodine, C. F. (1975). Interpreting chest radiographs without visual search. *Radiology, 116*(3), 527-532.

Kurtz, K. J., & Gentner, D. (2013). Detecting anomalous features in complex stimuli: The role of structured comparison. *Journal of Experimental Psychology: Applied*, *19*(3), 219–232. https://doi.org/10.1037/a0034395

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: https://doi.org/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: https://doi.org/10.3758/bf03196825

McCarley, J. S., Leggett, N., & Enright, A. (2020). Shared Gaze Fails to Improve Team Visual Monitoring. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 001872082090234. https://doi.org/10.1177/0018720820902347

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.

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: https://doi.org/10.1016/j.jmp.2008.02.001

Nakashima, R., Watanabe, C., Maeda, E., Yoshikawa, T., Matsuda, I., Miki, S., & Yokosawa, K. (2015). The effect of expert knowledge on medical search: medical experts have specialized abilities for detecting serious lesions. *Psychological Research*, *79*(5), 729–738. https://doi.org/10.1007/s00426-014-0616-y

Niehorster, D. C., Cornelissen, T., Holmqvist, K., & Hooge, I. (2019). Searching with and against each other: Spatiotemporal coordination of visual search behaviour in collaborative and competitive settings. *Attention, Perception, & Psychophysics*, *81*, 666-683. https://doi.org/10.3758/s13414-018-01640-0.

Nodine, C. F., Lauver, S. C., & Toto, L. C. (1996). Nature of Expertise in Searching Mammograms for Breast Masses. *Academic Radiology*, *3*, 1000–1006.

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: https://doi.org/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 Exprimental Psychology, 73,* 28-38.

Pratte, M. S. (2018). hbmem: Hierarchical Bayesian Analysis of Recognition Memory. R package version 0.3-3. https://CRAN.R-project.org/package=hbmem

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

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Reed, A. V. (1973). Speed-accuracy trade-off in recognition memory. *Science, 181*, 574-576. https://doi.org/10.1126/science.181.4099.574

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: https://doi.org/10.1080/17470210902853530

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. doi: https://doi.org/10.1177/1745691617708630.

Smith, J. D., Redford, J. S. Gent, L. C., & Washburn, D. A. (2005). Visual Search and the Collapse of Categorisation. *Journal of Experimental Psychology: General, 134*, 443-460. doi: https://doi.org/10.1037/0096-3445.134.4.443.

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

Sorkin, R. D., West, R., & Robinson, D. E. (1998). Group performance depends on the majority rule. *Psychological Science, 9*(6), 456-463

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

Sowden, P. T., Davies, I. R. L., & Roling, R. (2000). Perceptual learning of the detection of features in X-ray images: A functional role for improvements in adults' visual sensitivity. *Journal of Experimental Psychology: Human Perception & Performance, 26*(1), 379-390.

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.

Wahn, B., Czeszumski, A., Labusch, M., Kingstone, A., & König, P. (2020). Dyadic and triadic search: Benefits, costs, and predictors of group performance. *Attention, Perception, & Psychophysics* https://doi.org/10.3758/s13414-019-01915-0

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica, 41*, 67-85. https://doi.org/10.1016/0001-6918(77)90012-9

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis.Springer-Verlag New York.

Wolfe, J. (1994). Guided search 2.0 a revised model of visual search. *Psychonomic Bulletin & Review, 1,* 202-238.

Yamani, Y., Neider, M. B., Kramer, A. F., McCarley, J. S. (2017). Characterizing the Efficiency of Collaborative Visual Search With Systems Factorial Technology. *Archives of Scientific Psychology*, *5*, 1-9. https://doi.org/10.1037/arc0000030.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.