The wisdom of crowds for visual search

Mordechai Z. Juni[a,1] and Miguel P. Eckstein[a,b]

[a]Department of Psychological and Brain Sciences, University of California, Santa Barbara, CA 93106-9660; and [b]Institute for Collaborative Biotechnologies, University of California, Santa Barbara, CA 93106-5100

Edited by Jeremy M. Wolfe, Brigham & Women's Hospital/Harvard Medical School, Cambridge, MA, and accepted by Editorial Board Member Susan T. Fiske March 30, 2017 (received for review July 14, 2016)

Decision-making accuracy typically increases through collective integration of people's judgments into group decisions, a phenomenon known as the wisdom of crowds. For simple perceptual laboratory tasks, classic signal detection theory specifies the upper limit for collective integration benefits obtained by weighted averaging of people's confidences, and simple majority voting can often approximate that limit. Life-critical perceptual decisions often involve searching large image data (e.g., medical, security, and aerial imagery), but the expected benefits and merits of using different pooling algorithms are unknown for such tasks. Here, we show that expected pooling benefits are significantly greater for visual search than for single-location perceptual tasks and the prediction given by classic signal detection theory. In addition, we show that simple majority voting obtains inferior accuracy benefits for visual search relative to averaging and weighted averaging of observers' confidences. Analysis of gaze behavior across observers suggests that the greater collective integration benefits for visual search arise from an interaction between the foveated properties of the human visual system (high foveal acuity and low peripheral acuity) and observers' nonexhaustive search patterns, and can be predicted by an extended signal detection theory framework with trial to trial sampling from a varying mixture of high and low target detectabilities across observers (SDT-MIX). These findings advance our theoretical understanding of how to predict and enhance the wisdom of crowds for real world search tasks and could apply more generally to any decision-making task for which the minority of group members with high expertise varies from decision to decision.

group decision rules | signal detection theory | ideal observer analyses | wisdom of crowds

Groups of insects (1–4), fish (5–7), birds (8–10), mammals (11–14), and primates (15–18) have been shown to aggregate their individual judgments into group decisions for various tasks (19, 20). Although some groups seem to have leaders who make decisions alone on behalf of their groups (17, 21–23), it is difficult for individuals to outperform even simple aggregations of the entire group's individual judgments (4, 7, 9, 10, 19, 24–26). Perhaps that is why humans often make important decisions as a group (27–29), even if the only expedient (30, 31) but effective (24, 31–34) group decision mechanism is to use the simple majority voting rule (35).

Previous human studies have shown that combining people's judgments into group decisions can lead to accuracy benefits in various domains, such as estimation (36–38), detection (34, 39–44), identification (45–47), and prediction (46, 48–52), a phenomenon known as the wisdom of crowds (53). For artificial tasks, where perceptual decisions are limited only by noise that is internal to each observer's brain (i.e., no external noise), the maximum wisdom of crowd benefits are specified by the idealized signal detection theory model that treats observers' internal judgments as normally distributed and statistically independent (SDT-IND) (54). Such idealized environments are uncommon in real world perceptual tasks for which harnessing the wisdom of the crowds is of potential high interest.

Perceptual decisions for real world images, like the aerial and medical images shown in Fig. 1 *A* and *B*, are often limited by properties inherent to the images (56). The target of interest might be at a vantage point that makes it difficult to notice, other objects can occlude the target, and noise in the imaging process can reduce the target's detectability. All observers viewing the same images share the same external sources of variability, which lead to correlations in their judgments and reduce collective integration benefits (39, 57). Fig. 2 shows a theoretical example of how the benefits of optimal pooling specified by classic signal detection theory decrease as the correlation between observers' judgments increases (41, 58, 59).

Aside from reduced detectability of targets caused by external noise, real world perceptual tasks often include spatial uncertainty, requiring observers to scrutinize large visual areas for potential targets that might not be very visible in the visual periphery. Spatial uncertainty is the case for many life-critical tasks, such as doctors searching for lesions in medical images or intelligence analysts searching for particular objects in satellite and aerial images. How the search component of perceptual tasks affects the collective integration benefits is, however, unknown. No theoretical framework within signal detection theory has been developed to predict the wisdom of crowd benefits for visual search tasks.

Here, we explored and modeled the collective integration benefits for a search task (Fig. 1*D*) compared with a single-location task (Fig. 1*C*) and the prediction given by SDT-IND, which marks the maximum idealized collective integration benefits for statistically independent observers in single-location perceptual tasks (54). We also investigated whether simple majority voting is as effective for visual search as it is for typical simple perceptual tasks without spatial uncertainty, like our single-location task that we ran as a control, where it can approximate the benefits of optimal pooling (33, 34).

For each task, we evaluated the benefits of combining observers' decisions using commonly investigated pooling algorithms. We also recorded observers' eye movements to understand how variations in gaze behavior across observers during visual search might impact the collective integration benefits. Fixation statistics

Significance

Simple majority voting is a widespread, effective mechanism to exploit the wisdom of crowds. We explored scenarios where, from decision to decision, a varying minority of group members often has increased information relative to the majority of the group. We show how this happens for visual search with large image data and how the resulting pooling benefits are greater than previously thought based on simpler perceptual tasks. Furthermore, we show how simple majority voting obtains inferior benefits for such scenarios relative to averaging people's confidences. These findings could apply to life-critical medical and geospatial imaging decisions that require searching large data volumes and, more generally, to any decision-making task for which the minority of group members with high expertise varies across decisions.

Author contributions: M.Z.J. and M.P.E. designed research; M.Z.J. performed research; M.Z.J. and M.P.E. analyzed data; and M.Z.J. and M.P.E. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. J.M.W. is a guest editor invited by the Editorial Board.

[1]To whom correspondence should be addressed. Email: mzjuni@gmail.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1610732114/-/DCSupplemental.

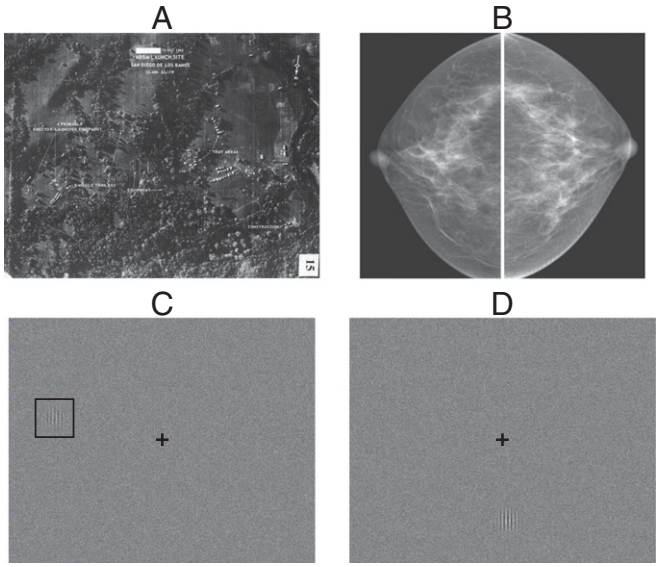

Fig. 1. Examples of perceptual tasks. (*A*) Real world example of an aerial image. Analysts from the National Photographic Interpretation Center determined that there were medium-range ballistic missiles in this reconnaissance photo that set off the Cuban missile crisis. Image courtesy of National Museum of the US Air Force. (*B*) Real world example of medical images. Radiologists would have to determine if there are cancerous lesions in these mammograms. Reprinted with permission from ref. 55; https://creativecommons.org/licenses/by-nc/3.0/). (*C*) Schematic of our single-location task. Each observer responded on an eight-point confidence scale whether there was a Gabor luminance patch (SNR = 3.5) in the middle of the black box. (*D*) Schematic of our search task. Each observer responded on an eight-point confidence scale whether there was a Gabor luminance patch (SNR = 7.88) anywhere in the image.

across observers were used to develop an extended signal detection theory framework with trial to trial sampling from a varying mixture of high and low target detectabilities across observers (SDT-MIX) that can predict the wisdom of crowd benefits for visual search.

Results

Twenty observers participated individually in the search task (Fig. 1*D*) followed by the single-location task (Fig. 1*C*) on a different day. Each task consisted of a yes–no signal detection task using noise-limited images. Each image (27.72° × 22.33°) was presented for 2,000 ms, and observers were informed that the target (a Gabor luminance patch) would be present in 50% of the images. The strength of the target was greater in the search task [signal to noise ratio (SNR) = 7.88] than in the single-location task (SNR = 3.5) to approximately equate observers' mean individual performance across tasks. Observers responded using an eight-point confidence scale, where one was highest confidence no, four was lowest confidence no, five was lowest confidence yes, and eight was highest confidence yes. This response mechanism elicited a binary yes–no decision together with a confidence rating for signal presence.

Individual Performance. Proportion correct (PC) is the proportion of trials in which the observer's binary yes–no decision was correct. Individual PCs ranged between 0.60 and 0.80 (M = 0.72, SD = 0.05) for the single-location task, ranged between 0.54 and 0.81 (M = 0.70, SD = 0.07) for the search task, and were not significantly different across tasks: t (19) = 1.66, $P > 0.05$ (paired samples t test). Each observer's index of detectability (d') was estimated using the area under the receiver operating characteristic (A_{ROC}) that emerges from his or her confidence ratings as follows (54, 60, 61): $d' = \sqrt{2}\,\Phi^{-1}(A_{ROC})$, where Φ^{-1} is the inverse cumulative distribution function of the standard normal distribution. Individual d' values ranged between 0.47 and 1.49 (M = 1.02, SD = 0.25) (Table S1) for the single-location task, ranged between 0.11 and 1.60 (M = 0.93, SD = 0.34) (Table S1) for the search task, and were not significantly different across tasks: t (19) = 1.22, $P > 0.05$ (paired samples t test).

Pooling Algorithms and Signal Detection Theory Predictions. We generated 500 random groups per group size and computed, using the same random groups for each task, the expected group performance of three commonly investigated pooling algorithms: averaging (AVG), weighted averaging (WAVG), and simple majority voting (MAJ). These algorithms were compared with the expected performance of the mean individual observer in the group (OBS) and the expected group performance predicted by SDT-IND or by classic signal detection theory for observer internal responses that are normally distributed and partially correlated (SDT-CORR).

OBS. For any given group, PC_{OBS} and d'_{OBS} are the mean individual PC and mean individual d' of the group, respectively.

MAJ algorithm. For any given group with an odd number of observers, the binary group decision of MAJ is no when the majority of the group's individual confidence ratings are between one and four and yes when the majority of the group's individual confidence ratings are between five and eight. Expected PC_{MAJ} and d'_{MAJ} for each group size were computed using the MAJ algorithm's binary group decisions (*Materials and Methods*).

AVG algorithm. For any given group and trial, the AVG algorithm compares the group's average confidence rating ($\hat{x}$) with a group decision criterion (c) (*SI Materials and Methods*). The binary group decision of AVG is no when $\hat{x} < c$ and yes when $\hat{x} > c$. Expected PC_{AVG} and d'_{AVG} for each group size were computed using the AVG algorithm's binary group decisions (*Materials and Methods*).

WAVG algorithm. The WAVG algorithm compares the group's weighted average confidence rating ($\hat{x} = \sum w_i x_i$) with a group decision criterion (c), where subscript i represents the ith group member. For any given group and trial, and using a leave one trial out procedure, we assigned weights (w) taking into account the covariance between the group's individual confidence ratings and how well each group member discriminates between signal and noise trials (*SI Materials and Methods*). The binary group decision of WAVG is no when $\hat{x} < c$ and yes when $\hat{x} > c$. Expected PC_{WAVG} and d'_{WAVG} for each group size were computed using the WAVG algorithm's binary group decisions (*Materials and Methods*).

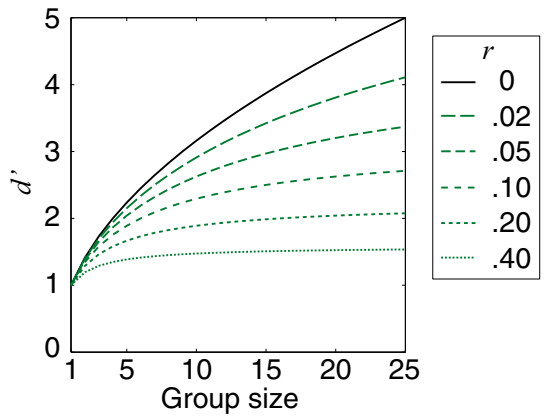

Fig. 2. Benefits of optimal pooling specified by classic signal detection theory. In this theoretical example, observers' internal judgments are normally distributed, and their individual d' values = 1. The black curve shows the d' obtained from optimal pooling as a function of group size when observers' internal judgments are statistically independent (see SDT-IND). The dashed green curves show the same when observers' internal judgments are correlated (see SDT-CORR). The d' obtained from optimal pooling decreases as the correlation (r) between observers' internal judgments increases.

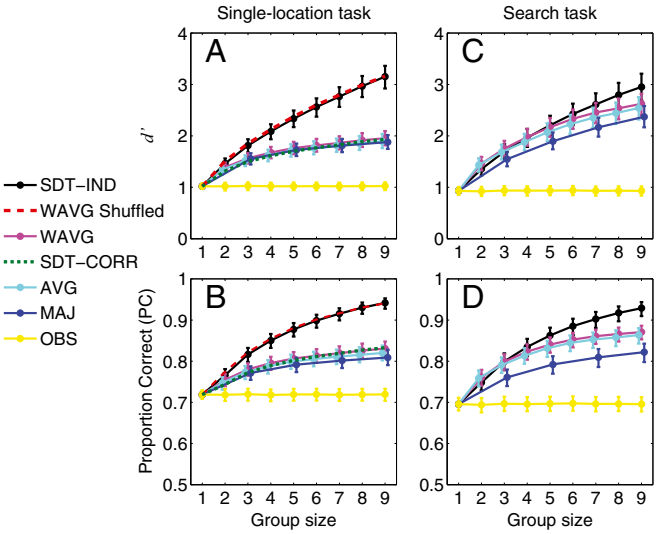

Fig. 3. Performance. (*A* and *C*) The *d′* values and (*B* and *D*) the PCs of each pooling algorithm (AVG, WAVG, and MAJ), classic signal detection theory prediction (SDT-IND and SDT-CORR), and OBS are plotted as a function of group size for each task. The text has details regarding WAVG Shuffled and why we do not show WAVG Shuffled or SDT-CORR for (*C* and *D*) the search task. Data points for group size = 1 mark mean individual performance (n = 20), and error bars mark ±SEM. Data points for all other group sizes were computed based on 500 random groups per group size using the same random groups for both tasks, and error bars mark bootstrap 68.27% confidence intervals to be equivalent to the percentile of ±SD of a normal distribution.

SDT-IND and SDT-CORR predictions. For any given group of m observers, the group performances predicted by SDT-IND and SDT-CORR are computed as follows (54, 59): $d'_{\text{SDT-IND}} = \sqrt{\sum (d'_i)^2}$, where subscript i represents the ith group member, and $d'_{\text{SDT-CORR}} = \sqrt{Var(d')m/(1-r) + (\overline{d'})^2 m/(1+(m-1)r)}$, where $\overline{d'}$ and $Var(d')$ are the mean and variance of the group's individual d' values, and r represents the correlation between the group's individual confidence ratings. Given each predicted d'_{SDT}, we computed the $\text{PC}_{\text{SDT-IND}}$ that would be predicted for an optimally placed criterion at $d'_{\text{SDT-IND}}/2$ and the $\text{PC}_{\text{SDT-CORR}}$ that would be predicted for an optimally placed criterion at $d'_{\text{SDT-CORR}}/2$ as follows (54, 61): $\text{PC}_{\text{SDT}} = \Phi(d'_{\text{SDT}}/2)$, where Φ is the cumulative distribution function of the standard normal distribution. [Note that the PC predicted for an optimal criterion placed halfway between equal variance signal and noise distributions is $\Phi(d'/2)$ for yes–no tasks that have 50% signal prevalence vs. $\Phi(d'/\sqrt{2})$ for two-alternative forced choice tasks, although both kinds of tasks have the same optimal criterion at $d'/2$.]

Group Performance. Fig. 3 shows expected d' and PC as a function of group size for each task. Comparison of d' values and PCs across algorithms and tasks suggests (*i*) that WAVG and AVG achieve closer group performance to the SDT-IND prediction in the search task than in the single-location task and (*ii*) that MAJ underperforms WAVG and AVG in the search task but not in the single-location task. To evaluate algorithms and tasks controlling for residual differences in task difficulty and individual performance, we computed relative efficiency metrics as defined below.

Observer efficiency is a well-known metric that quantifies a human observer's detection sensitivity with respect to an ideal detector by taking the squared ratio between human and ideal contrast thresholds or d' values (62, 63). Similarly, group efficiency has been used to quantify group performance with respect to an ideal detector (64) or ideal group performance (25) by taking the squared ratio between the relevant contrast thresholds or d' values. Here, we are interested in comparing the group performance of each algorithm relative to the other algorithms, predictions, and mean individual performance. Hence, we define an algorithm's relative efficiency as the squared ratio between the d' of the algorithm in question and the d' of the comparison of interest (65, 66). For example, the efficiency of MAJ relative to WAVG is equivalent to $(d'_{\text{MAJ}}/d'_{\text{WAVG}})^2$. [Comparisons using PC ratios are problematic because of the nonlinear compressive relationship between PC and d' (e.g., doubling or tripling the SNR of a stimulus does not double or triple the consequent PC). This nonlinear relationship means that a constant source of suboptimality across tasks, such as incomplete integration of information across observers, would lead to a constant d' ratio but would not lead to a constant PC ratio (65).]

Fig. 4 shows the expected efficiency of each algorithm as a function of group size for each task relative to the mean individual OBS in the group (Fig. 4 *B* and *D*) and relative to the SDT-IND prediction (Fig. 4 *A* and *C*). Relative efficiency > 1 indicates that the algorithm in question achieves higher performance than the comparison of interest. Relative efficiency < 1 indicates that the algorithm in question achieves lower performance than the comparison of interest. To follow, we report the statistical analyses of the relative efficiency metrics using bootstrap resampling methods (*SI Materials and Methods*) (67).

Comparison between algorithms within each task.

Single-location task. Fig. 4 *A* and *B* shows little difference in relative efficiency for the single-location task between the AVG, WAVG, and MAJ algorithms. The efficiency of each algorithm relative to the mean individual OBS in the group (Fig. 4*B*) is significantly higher than one for all group sizes that we tested (all P values < 0.001; bootstrap resampling), and the efficiency of each algorithm relative to SDT-IND (Fig. 4*A*) is significantly lower than one for all group sizes that we tested (all P values < 0.005; bootstrap resampling). A direct comparison between algorithms shows that the efficiency of MAJ relative to AVG (ranging between 0.99 and 1.05) and relative to WAVG (ranging between 0.92 and 0.99) is not significantly different from one for any group size that we tested (vs. AVG: all P values > 0.39; vs. WAVG: 0.13 > all P values ≥ 0.05; bootstrap resampling), which indicates

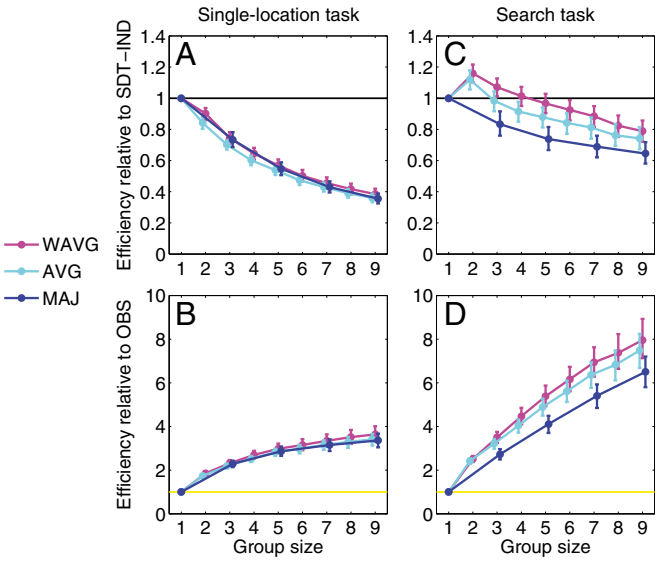

Fig. 4. Relative efficiency. The efficiency of each pooling algorithm is plotted as a function of group size for each task relative to (*A* and *C*) the SDT-IND prediction and (*B* and *D*) the mean individual OBS in the group. Note that the algorithms have similar relative efficiency for (*A* and *B*) the single-location task but that AVG and WAVG have higher relative efficiency than MAJ for (*C* and *D*) the search task. Also note that each algorithm's relative efficiency is generally higher for (*C* and *D*) the search task than for (*A* and *B*) the single-location task. Fig. 3 has a summary of how we computed data points and error bars.

that the benefits obtained by MAJ are similar to AVG and can approximate the benefits obtained by WAVG in this task.

The underperformance of collective integration relative to the SDT-IND prediction in this task is consistent with the correlation between observers' judgments caused by the external noise in each image that was common for all observers. If the classic signal detection theory prediction includes the mean estimated correlation between observers' confidence ratings ($r = 0.26$), then the prediction for correlated observers, SDT-CORR, closely matches WAVG as shown in Fig. 3 *A* and *B*. Furthermore, if we shuffle the trials across observers to artificially remove the correlation between observers' confidence ratings, then the performance of weighted averaging for shuffled trials (WAVG Shuffled) should correspond to the idealized signal detection theory prediction for statistically independent observers. Indeed, Fig. 3 *A* and *B* shows that WAVG Shuffled closely matches SDT-IND. (We do not show SDT-CORR or WAVG Shuffled for the search task, because they do not match WAVG and SDT-IND, respectively, in that task such as they do in the single-location task, which suggests the inadequacy of using classic signal detection theory to model the collective integration benefits of the search task.)

These results for the single-location task serve to show (*i*) that classic signal detection theory is a valid predictive model for the collective integration benefits of this simple perceptual task and (*ii*) that simple majority voting can approximate the benefits of weighted averaging and the upper limit set by classic signal detection theory for this simple perceptual task.

Search task. Fig. 4 *C* and *D* shows considerable difference in relative efficiency for the search task between the MAJ algorithm and the AVG and WAVG algorithms. Although the efficiency of each algorithm relative to the mean individual OBS in the group (Fig. 4*D*) is significantly higher than one for all group sizes that we tested (all P values < 0.001; bootstrap resampling), direct comparison between algorithms shows that the efficiency of MAJ relative to AVG (ranging between 0.84 and 0.87) and relative to WAVG (ranging between 0.76 and 0.82) is significantly lower than one for all group sizes that we tested (vs. AVG: all P values < 0.03; vs. WAVG: all P values < 0.002; bootstrap resampling). Thus, although all algorithms achieve collective integration benefits, MAJ underperforms both AVG and WAVG in this task.

In addition, although the efficiency of MAJ relative to SDT-IND (Fig. 4*C*) is significantly lower than one for all group sizes that we tested (all P values < 0.001; bootstrap resampling), the efficiency of WAVG relative to SDT-IND is significantly higher than one for groups of two ($P < 0.03$; bootstrap resampling) and is not significantly different from one for groups of three and four [not significant (ns); bootstrap resampling] but is significantly lower than one for groups of five through nine (all P values < 0.05; bootstrap resampling). Similarly, the efficiency of AVG relative to SDT-IND is not significantly different from one for groups of two and three (ns; bootstrap resampling) but is significantly lower than one for groups of four through nine (all P values < 0.006; bootstrap resampling). Thus, although MAJ underperforms the SDT-IND prediction for all group sizes in this task, AVG and WAVG can approach the prediction of SDT-IND in this task, at least for small group sizes.

These results for the search task show (*i*) that averaging and weighted averaging can approach the prediction of SDT-IND for small group sizes in this task (unlike in the single-location task, where the external noise in our images induces correlations between observers' judgments, which guarantees that no pooling algorithm can approach the idealized prediction of SDT-IND for any group size in that task) and (*ii*) that simple majority voting does not approximate the higher benefits of averaging and weighted averaging in this task. In *Explaining why MAJ underperforms AVG and WAVG in the search task*, we explain why simple majority voting underperforms averaging and weighted averaging in the search task.

Comparison across tasks. Comparison of the algorithms' higher relative efficiencies for the search task (Fig. 4 *C* and *D*) with the algorithms' lower relative efficiencies for the single-location task (Fig. 4 *A* and *B*) suggests that collective integration benefits are greater for the search task than for the single-location task.

The efficiency of AVG and WAVG relative to the SDT-IND prediction (Fig. 4 *A* and *C*) and relative to the mean individual OBS in the group (Fig. 4 *B* and *D*) is significantly higher in the search task (Fig. 4 *C* and *D*) than in the single-location task (Fig. 4 *A* and *B*) for all group sizes that we tested (all P values < 0.001; bootstrap resampling). These results indicate that the benefits of AVG and WAVG are unequivocally greater for the search task than for the single-location task. The efficiency of MAJ relative to the SDT-IND prediction (Fig. 4 *A* and *C*) and relative to the mean individual OBS in the group (Fig. 4 *B* and *D*) is significantly higher in the search task (Fig. 4 *C* and *D*) than in the single-location task (Fig. 4 *A* and *B*) for groups of five (all P values < 0.05), seven (all P values < 0.02), and nine (all P values < 0.01) but is not significantly different across tasks for groups of three (ns; bootstrap resampling). These results indicate that, for a subset of group sizes, the benefits of MAJ can be higher for the search task than for the single-location task (but not the other way around). Finally, the efficiency of MAJ relative to AVG and WAVG is significantly lower in the search task (ranging between 0.76 and 0.87) than in the single-location task (ranging between 0.92 and 1.05) for groups of three (all P values < 0.005), five (all P values < 0.02), and seven (all P values < 0.04) but is not significantly different across tasks for groups of nine (ns; bootstrap resampling).

These results for the comparison of relative efficiencies across tasks indicate that averaging, weighted averaging, and to some extent, simple majority voting can attain greater benefits in the search task than in the single-location task. In *Explaining why the search task obtains greater collective integration benefits relative to the single-location task*, we explain why the search task obtains greater collective integration benefits relative to the single-location task.

Analysis of Gaze Behavior During Visual Search. We hypothesize that the dissociations across algorithms and tasks described above arise from an interaction between the foveated properties of the human visual system (high acuity at fixation and low acuity in the periphery) and observers' nonexhaustive search patterns. Because the target is of medium spatial frequency (six cycles per 1°), its detectability degrades rapidly with retinal eccentricity (68). This low target detectability using peripheral vision plays a role in the search task, where the target could be viewed foveally or peripherally, as opposed to the single-location task, where the target is always viewed foveally.

For our particular search task, fixating near the target increases the probability of detecting it as shown by observers' individual hit rates, which were significantly higher when they foveated near the target ($M = 0.88$, SD = 0.08) than when they did not [$M = 0.32$, SD = 0.09, t (19) = 22.94, $P < 0.0001$ (paired samples t test)] (leftmost data points in Fig. 5*B*). Observers in our task averaged 5.7 fixations per trial and did not foveate near the target on every trial (Fig. 5*C* shows examples of fixation paths). The percentage of signal trials that each observer foveated near the target ranged between 23.6 and 69.6% ($M = 49.6\%$, SD = 12.4%) (Table S1), and the pairwise correlation between observers as to whether they foveated near the target ranged between −0.02 and 0.43 ($M = 0.21$, SD = 0.08) (Fig. S1). We reasoned that, if variations in gaze behavior (and consequent effects on target detectability) give rise to the greater collective integration benefits for visual search, then we should see different levels of collective integration benefits in the search task for different patterns of gaze behavior across group members.

Using observers' gaze position data, we partitioned the signal trials of the search task into four categories based on the proportion (p) of observers in the group who foveated within 2° from the center of the target: all (p = 1), a majority ($0.5 < p < 1$), a minority ($0 < p < 0.5$), or none (p = 0). For reference, Fig. 5*A* shows each algorithm's hit rate across all groups and trials, irrespective of the proportion of observers in the group who foveated near the target. Fig. 5*B* shows that, when all observers

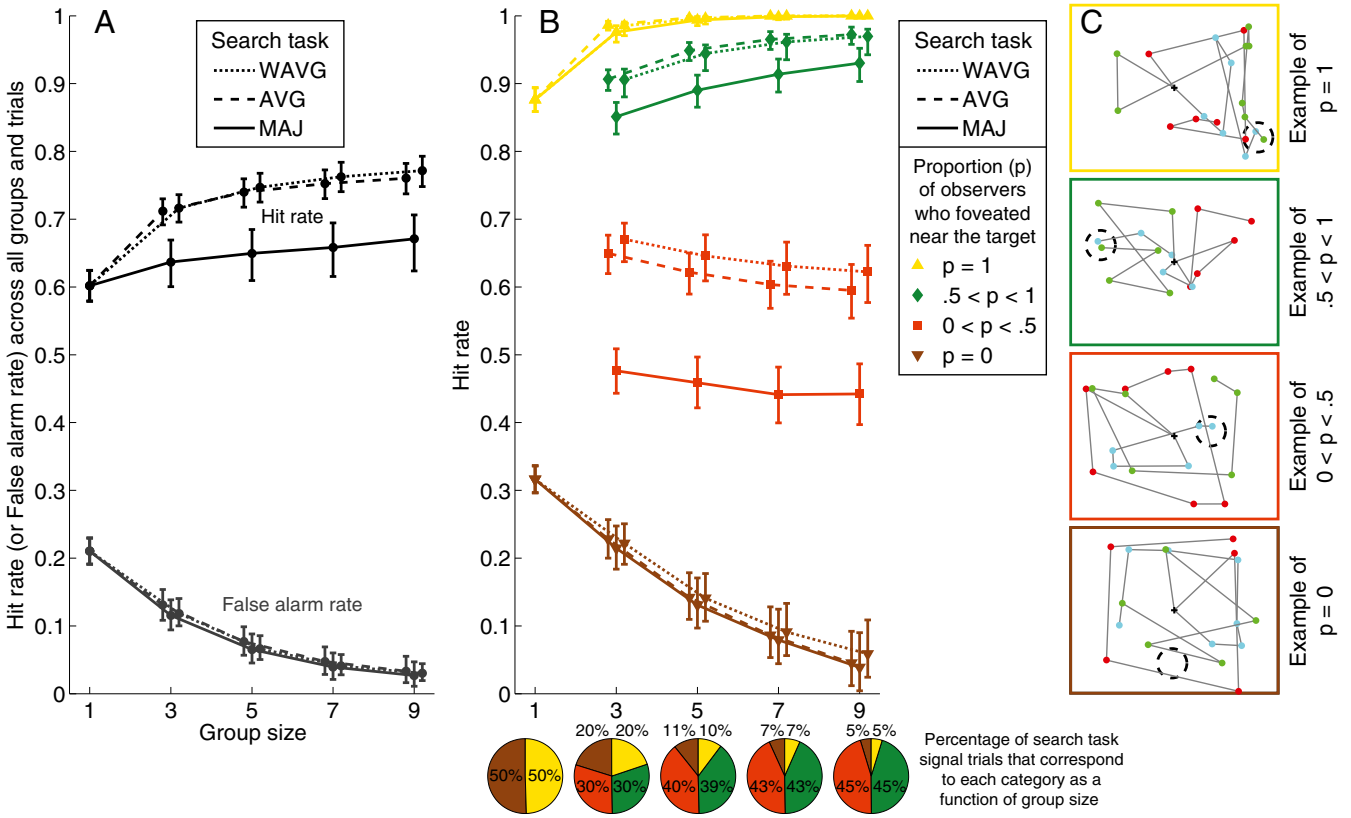

Fig. 5. Hit rate breakdown for the search task. (*A*) Overall hit rate (across all signal trials) and overall false alarm rate (across all noise trials) of each algorithm are plotted as a function of group size for the search task. (*B*) Each algorithm's hit rate is broken down for each group size as a function of the proportion (p) of observers in the group who foveated near the target: all (p = 1), a majority (0.5 < p < 1), a minority (0 < p < 0.5), or none (p = 0). The pie charts show the percentage of signal trials across all 500 random groups that belong to each category for each group size. (*C*) Schematics of actual gaze paths for three observers during four different trials. From top to bottom, the panels show respective examples when all, a majority, a minority, and none of the observers in the group foveated within 2° from the center of the target. Fig. 3 has a summary of how we computed data points and error bars.

in the group foveated near the target (p = 1), the hit rate was close to one for all algorithms regardless of group size, and the resulting efficiency relative to SDT-IND was higher than 3.30 for all algorithms regardless of group size (Fig. S3). Conversely and as shown in Fig. 5*B*, when none of the observers in the group foveated near the target (p = 0), the hit rate ranged between 0.04 and 0.23 depending on the algorithm and group size, and the efficiency relative to SDT-IND was close to 0 for all algorithms regardless of group size (Fig. S3).

Most signal trials were ones in which a majority (but not all) or minority of observers in the group foveated near the target. As shown in Fig. 5*B*, when a majority of observers in the group foveated near the target (0.5 < p < 1), the hit rate ranged between 0.85 and 0.97 depending on the algorithm and group size, and the resulting efficiency relative to SDT-IND ranged between 1.33 and 2.17 depending on the algorithm and group size (Fig. S3). Finally and as shown in Fig. 5*B*, when a minority of observers in the group foveated near the target (0 < p < 0.5), the hit rate ranged between 0.44 and 0.67 depending on the algorithm and group size, and the efficiency relative to SDT-IND ranged between 0.36 and 0.92 depending on the algorithm and group size (Fig. S3).

The aggregate collective integration benefits for the search task are a combination of the benefits across all of the different trial types mentioned above, which include a large percentage of signal trials (50% regardless of group size) that have very high benefits when all or a majority of observers in the group foveated near the target. These trials more than compensate for the smaller percentage of signal trials (5–20% depending on group size) that do not have any benefits (efficiency relative to SDT-IND ~0) when none of the observers in the group foveated near the target. Additionally, although the remaining percentage of signal trials (30–45% depending on group size) has modest benefits when a minority of observers in the group foveated near the target, the efficiency for those trials relative to SDT-IND, which ranged between 0.36 and 0.92 depending on the algorithm and group size (Fig. S3), is comparable with that in the single-location task, where efficiency relative to SDT-IND for the same group sizes ranged between 0.36 and 0.73 depending on the algorithm and group size (Fig. 4*A*).

Explaining why the search task obtains greater collective integration benefits relative to the single-location task. The breakdown of benefits described above, which in the aggregate, exceed the collective integration benefits for the single-location task, is an emergent statistical property of the search task, where observer confidence ratings for each signal trial are sampled from a varying mixture of high and low target detectabilities across observers: high detectability for those who viewed the target foveally on that trial (mean $d' = 1.96$) (Table S1) and low detectability for those who only viewed the target peripherally on that trial (mean $d' = 0.25$) (Table S1). Critically, our observers were not perfectly correlated as to whether they foveated near the target (mean pairwise correlation = 0.21) (Fig. S1). Thus, even worse-performing group members could contribute to increasing the algorithm's group performance, because they sometimes happen to foveate near the target on trials that some of the better-performing group members do not.

For example, if we remove the two worst-performing individuals from each group of five to create artificial groups of three, then the respective d' values of AVG, WAVG, and MAJ decrease from 2.07, 2.17, and 1.90 for the original groups of five

(Fig. 3*C*) to 1.92, 2.00, and 1.78 for the artificial groups of three that omit the contributions of the two worst-performing individuals from each group (as measured by their individual d' values for the search task) (Table S1). The resulting efficiencies of these artificial groups of three relative to the original groups of five are 0.86, 0.85, and 0.88 for the AVG, WAVG, and MAJ algorithms, respectively. These relative efficiency values can be compared with an equivalent analysis for the single-location task, where the efficiencies of the artificial groups of three that omit the contributions of the two worst-performing individuals from each group (as measured by their individual d' values for the single-location task) (Table S1) relative to the original groups of five are 0.95, 0.93, and 0.97 for the AVG, WAVG, and MAJ algorithms, respectively. These relative efficiency values for the single-location task are closer to one and higher than for the search task, suggesting lower contributions from low-performing group members for the single-location task.

The intuition for why the search task obtains greater collective integration benefits is that many signal trials end up having a varying subset of observers who happen to view the target foveally with relatively high target detectability, because the target (peak contrast = 18%; SNR = 7.88) is easy to detect when using foveal vision (but very difficult to detect when using peripheral vision). This situation is in contrast to the single-location task, where the target is somewhat difficult to detect when using foveal vision (peak contrast = 8%; SNR = 3.5), but all observers always viewed it foveally, which means that the observer confidence ratings for each trial are sampled from a constant set of (medium) target detectabilities across observers. Additionally, although some targets in the single-location task might be easier to detect than others, that increased foveal detectability (because of randomly lower noise) is common for all observers. Hence, it is never the case that some individuals in the single-location task get to process the target for a particular trial with incidentally higher than usual target detectability whereas others do not. Therefore, although there are substantial wisdom of crowd benefits for the single-location task, those benefits are surpassed in the search task, because it engenders trial to trial sampling from a varying mixture of high and low target detectabilities across observers contingent on viewing the target foveally vs. peripherally. Also, as shown above, even low-performing group members could contribute to increasing the algorithm's group performance in the search task.

Explaining why MAJ underperforms AVG and WAVG in the search task. We showed that the efficiency of MAJ relative to AVG and relative to WAVG is significantly lower than one in the search task for all group sizes that we tested. The disadvantage of MAJ in this task cannot be explained in terms of WAVG's differential weighting parameter that generally assigns more weight to those individuals who foveated near the target more often and had higher d' values (Table S1), because AVG does not have this differential weighting parameter and still outperforms MAJ in this task.

Fig. 5*B* suggests that the disadvantage of MAJ in this task is driven in large part by the higher hit rate that AVG and WAVG achieve over MAJ when a minority of observers in the group foveated near the target ($0 < p < 0.5$). This hit rate divergence occurs, because some of those individuals in the minority of the group who happened to foveate near the target on a particular trial often had high-enough confidence ratings to push the group's average and weighted average ratings above AVG's and WAVG's respective group decision criterion, even if the majority of observers in the group who did not happen to foveate near the target on that particular trial said no (Fig. S4 shows the confidence rating frequencies for each task).

Fig. 6 illustrates this phenomenon by analyzing, for each task, the distribution of confidence ratings across group members for the subset of trials in which the AVG algorithm reaches a yes decision, whereas the MAJ algorithm reaches a no decision. Using groups of seven as an example, Fig. 6 shows that, for this subset of trials, there were many more signal trials in the search task (green bars in Fig. 6, *Right*) than in the single-location task (green bars in Fig. 6, *Left*) for which there were one or more individuals in the group who chose the highest confidence rating for signal presence (Fig. S5 shows similar results comparing WAVG with MAJ).

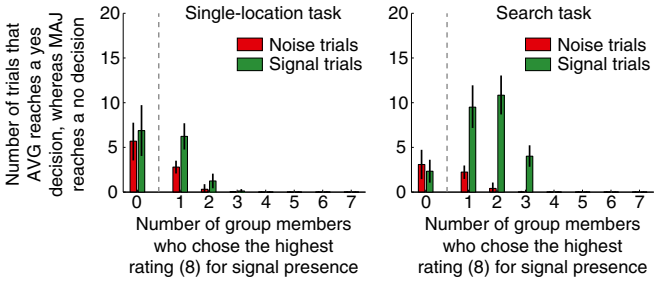

Fig. 6. Subset of trials in which the AVG algorithm reaches a yes decision, whereas the MAJ algorithm reaches a no decision. Histogram for the number of individuals in the group who chose the highest confidence rating for signal presence across trials in which the AVG algorithm reaches a yes decision, whereas the MAJ algorithm reaches a no decision. The histograms show the average of 500 random groups using the same random groups for both tasks, and the error bars mark bootstrap 68.27% confidence intervals. This figure shows results for groups of seven, and the results are similar for groups of three, five, and nine. For this subset of trials, there were many more signal trials in the search task than in the single-location task for which there was one or more observers in the group who chose the highest rating for signal presence (eight on the eight-point confidence scale). Fig. S5 shows similar results comparing WAVG with MAJ.

These results show that only the search task, which engenders a trial to trial sampling from a varying mixture of high and low target detectabilities across observers, obtains a considerable number of signal trials where a random majority of observers in the group incorrectly says that the target was absent (because they only processed the target with low peripheral detectability on that particular trial), whereas the remaining minority of observers in the group correctly says that the target was present with very high confidence (because they happened to process the target with high foveal detectability on that particular trial). And unlike the majority voting procedure that does not make use of how confident each observer is, the averaging procedure (whether weighted or not) is able to exploit the very high confidence that is often provided by those individuals in the group who happened to process the target with high foveal detectability on that particular trial (green bars in Fig. S4, *Right*).

Extended Signal Detection Theory Framework. We hypothesized above that the greater collective integration benefits for visual search and the higher benefits of AVG and WAVG over MAJ for visual search arise from having a varying mixture of target detectabilities across observers, where each observer has high vs. low target detectability on each signal trial contingent on viewing the target foveally vs. peripherally. If this hypothesis for visual search is correct, then we should be able to implement an extended signal detection theory framework (SDT-MIX) that generates collective integration benefits that resemble the empirically measured benefits of AVG, WAVG, and MAJ in our search task.

Fig. 7, *Left* shows a sketch of the extended SDT-MIX framework applied to our search task. The framework assumes that, for each signal trial, each individual has q% chance of processing the target foveally with his or her high target detectability (high d') and 100 – q% chance of processing the target peripherally with his or her low target detectability (low d'). Not shown in the sketch but assumed by the SDT-MIX framework is that individuals can be correlated as to whether they process the target foveally vs. peripherally (Fig. S1).

We simulated SDT-MIX using the following range of input values for 20 theoretical individuals to resemble our empirical observers in the search task (Fig. S1 and Table S1, left side show detailed input values); 20 individual high d' values for foveal target processing ranged between 0.76 and 2.50, 20 individual

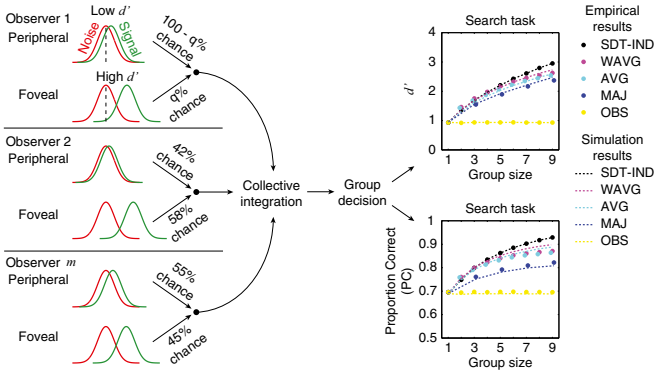

Fig. 7. Extended SDT-MIX framework and simulation results for the search task. *Left* depicts a sketch of the extended signal detection theory framework with a varying mixture of high and low target detectabilities across observers contingent on processing the target foveally vs. peripherally on each signal trial. The AVG, WAVG, and MAJ collective integration algorithms were applied to Gaussian random variables sampled from this framework. *Right* shows the simulation results of this framework applied to our search task (dashed lines) (Fig. S6 shows simulated relative efficiencies) together with the empirical results of our search task (circles) (Fig. 3 *C* and *D* shows empirical error bars).

low d' values for peripheral target processing ranged between –0.09 and 0.59, 20 individual q% chances on each signal trial of getting high d' instead of low d' ranged between 23.6 and 69.6%, and 190 pairwise correlations between 20 individuals in getting high d' vs. low d' ranged between –0.02 and 0.43. (*SI Materials and Methods* has additional simulation details, and Fig. S2*A* shows a theoretical example of how collective integration benefits predicted by the SDT-MIX framework would decrease with increased correlation between individuals in getting high d' vs. low d'.)

The m group members' individual judgments for each noise trial were sampled as Gaussian random variables from a common noise distribution for all individuals. However, the m group members' individual judgments for each signal trial were sampled as Gaussian random variables from one of 2^m possible combinations of signal distributions across individuals, depending on which individuals got high d' vs. low d' for that trial through correlated Bernoulli sampling processes (69).

Fig. 7, *Right* shows the simulation results of the SDT-MIX framework applied to our search task (dashed lines in Fig. 7, *Right*) together with the empirical results of our search task (circles in Fig. 7, *Right*). The close match (with no fitting parameters) between the simulated and empirical results of the AVG, WAVG, and MAJ algorithms suggests that the greater collective integration benefits for the search task and the higher benefits of AVG and WAVG over MAJ for the search task arise, as discussed above, from statistical principles related to the trial to trial sampling from a varying mixture of high and low target detectabilities across observers as modeled by SDT-MIX (as opposed to the trial to trial sampling from a constant set of target detectabilities across observers in the single-location task).

We note that the WAVG algorithm outperformed the SDT-IND prediction in our search task but only for groups of two where the empirical efficiency of WAVG relative to SDT-IND was significantly higher than one with $P < 0.03$ (bootstrap resampling). Although our simulation of the SDT-MIX framework did not obtain this qualitative result of WAVG outperforming SDT-IND for any group size (purple dashed line in Fig. S6, *Upper*), we show in *SI Materials and Methods* that WAVG could potentially outperform the SDT-IND prediction for many group sizes for circumstances with very low interobserver correlations in getting high d' vs. low d' (Fig. S2*A*). The SDT-MIX simulation applied to our search task used interobserver correlations in getting high d' vs. low d' to resemble the empirically estimated pairwise correlations between observers in foveating near the target (Fig. S1). Those estimates assume that high detectability processing occurs when observers foveate within 2° from the center of the target. Using a stricter visual angle threshold would reduce the estimated pairwise correlations between observers in foveating near the target, but we used a 2° threshold because of limitations with eye-tracking precision. Thus, a possible explanation for why the SDT-MIX simulation of our search task did not qualitatively obtain the empirical aspect of WAVG outperforming SDT-IND for groups of two might be that we overestimated how correlated observers were in processing the target foveally with high detectability vs. peripherally with low detectability.

Discussion

How to Predict the Wisdom of Crowd Benefits for Visual Search. We showed that, for single-location perceptual tasks, classic signal detection theory (54, 59) well-predicts collective integration benefits, because observer judgments for each image are elicited from a constant set of target detectabilities across observers (34). For visual search tasks, however, we empirically showed that classic signal detection theory is not a valid model to predict collective integration benefits. We hypothesized that the greater collective integration benefits for visual search arise from an interaction between the foveated nature of the human visual system and observers' nonexhaustive search patterns. This interaction would give rise to a situation in which, from trial to trial, a varying observer or subset of observers would fixate near the target and process it with high foveal detectability, whereas the remaining observers would process it with low peripheral detectability. This hypothesis was supported by the empirical variation in group performance with observers' gaze position data, and by the close match between the empirical pooling benefits for the search task and the simulated pooling benefits obtained using the STD-MIX framework.

Our theory makes specific predictions that could potentially be tested in future studies. For example, the dissociation in collective integration benefits between the single-location vs. search tasks should decrease for targets that are more visible in the visual periphery. For targets that are equally detectable across the visual field, the dissociation across tasks should vanish altogether, and classic signal detection theory should well-predict the collective integration benefits for both tasks. Similarly, if observers are given unlimited time and instructed to thoroughly fixate all image regions, then the dissociation in collective integration benefits across tasks should once again vanish.

Simple Majority Voting Obtains Inferior Wisdom of Crowd Benefits for Visual Search. We showed that, for the single-location task, simple majority voting can approximate the upper limit for collective integration benefits obtained by weighted averaging of observers' confidences as specified by classic signal detection theory (54, 59). For the search task, however, we empirically showed and theoretically modeled using the extended SDT-MIX framework that simple majority voting does not approximate the higher benefits of averaging and weighted averaging of observers' confidences.

This finding that simple majority voting obtains inferior accuracy benefits for visual search relative to averaging and weighted averaging could be relevant for real world search tasks, because human groups have a propensity to use a majority voting procedure to reach their joint decisions in various domains (27, 28, 31, 33, 34). For many real world perceptual tasks that involve a search component, it might be appropriate to abandon the easy majority voting procedure in favor of a more demanding averaging of confidences procedure to obtain higher collective integration benefits.

We previously showed that human groups in certain circumstances can infer that different observers might have access to different amounts of information at different times and that they can dynamically adapt their joint decision algorithms away from simple majority voting to improve their joint decision accuracies (34). Others have shown that human groups can derive benefits by altering their individual decision strategies when searching for targets collaboratively (70–72). Future studies exploring collective wisdom for real world search tasks can explore

whether groups adapt their joint decision algorithms to use more demanding averaging or weighted averaging procedures that obtain higher joint decision accuracies than the default majority voting procedure.

Applicability of Our Findings to Real World Search Tasks. Many studies have shown benefits in combining expert judgments and/or crowdsourcing nonexpert judgments for a variety of disciplines, including radiology (41, 42, 44), ophthalmology (73, 74), pathology (44, 75, 76), and clinical predictions (52). However, few studies (41, 42, 77) have compared the benefits of various pooling algorithms with those predicted by signal detection theory. Here, we experimentally showed a visual search task for which the collective integration benefits depart from classic signal detection theory predictions and are consistent instead with our extended SDT-MIX framework. The applicability of our findings to real world search tasks requires two conditions: (*i*) a detectability for the target that degrades rapidly with retinal eccentricity (78) and (*ii*) individuals not thoroughly fixating all image regions (79, 80).

In our search task, these conditions were guaranteed by using a limited viewing time (2,000 ms) and a midfrequency Gabor target (six cycles per 1°) that degrades rapidly in detectability with retinal eccentricity (68). Although studies with realistic targets in naturalistic scenes (81) and medical images (82) have shown steep degradations in detectability with retinal eccentricity (thus meeting the first condition), real world searches can have unlimited viewing time, allowing observers to thoroughly fixate all image regions if desired. This discrepancy in viewing time might bring into question the applicability of our findings to real world search tasks, and we agree that thoroughly fixating all image regions is possible, if desired, for planar medical imaging or baggage screening (which consist of one X-ray or a small number of X-rays per case), so that the benefits of collective integration could be modeled correctly using classic signal detection theory in agreement with previous studies (41, 42, 77).

However, many life-critical search tasks involve large datasets, preventing observers from realistically scrutinizing all image regions (79, 80). Radiologists are increasingly using 3D volumetric imaging consisting of a large number of X-ray slices per scan: 64–128 slices for computed tomography (83) and 50–90 slices for digital breast tomosynthesis (DBT) (84–86). Based on clinical reading times of 2–3 min per case (84, 85) and 250–350 ms per fixation, radiologists would average no more than 14.4 fixations per slice when inspecting one DBT scan (3 min of 250 ms per fixation across 50 slices for one breast) and possibly as little as 1.9 fixations per slice when inspecting a separate DBT scan for each breast (2 min of 350 ms per fixation across 180 total slices for both breasts). Similarly, the increasing amounts of surveillance photographs taken across space and time by unmanned aerial vehicles and ever-cheaper quadcopters might not permit geospatial analysts to thoroughly fixate all aerial image regions for a case.

Our approach involves combining observer confidence ratings about signal presence without any localization judgments. This focus on confidence ratings is often the case for double-reading breast cancer screening programs, for example, where localization judgments are not generally required to determine whether patients should be recalled for additional tests. Furthermore, many European and Australian screening centers use two independent readers for each case, with arbitration by a third reader when the first two are discordant. The final decision whether the patient should be recalled typically follows the majority voting rule without explicit consideration of any localization information. However, in many other applications, observers might be asked to mark the possible location of the target. Such localization judgments could potentially guide how to combine confidence ratings across observers to improve performance. Finally, in circumstances where observers have a second viewing of the images or a discussion, localization judgments from observers with high confidence that the target is present could lead others who originally missed the target to revise their response (87), thus influencing the relationship between different group decision algorithms and signal detection theory predictions for second-stage decisions.

Generalization to Other Decision-Making Tasks. The implications of our study and proposed SDT-MIX framework could apply more generally to any scenario where, from decision to decision, a varying individual or minority of individuals in the group often has substantially higher probability of reaching a correct decision compared with the remaining majority of individuals in the group (88). One example would involve a panel that is given a battery of questions spanning multiple domains and a scenario in which, from decision to decision, a varying minority of panelists often has high expertise and expresses very high confidence for their decision. Our extended SDT-MIX framework would suggest that collective integration benefits for this panel might be greater than otherwise expected from classic signal detection theory, and that a more demanding averaging of confidences might obtain significantly higher group performance than simple majority voting. Developing quantitative methods to identify scenarios that would substantially benefit from averaging people's confidences instead of simply following the majority decision could be an important objective for future research on how to improve collective integration of human decisions in different domains.

Materials and Methods

All data are available on request. *SI Materials and Methods* has additional materials and methods.

PC and d' for MAJ, AVG, and WAVG. Given each algorithm's binary group decisions across all groups and trials, we computed the algorithm's hit rate (proportion of signal trials across all groups in which the algorithm's binary group decision is yes) and false alarm rate (proportion of noise trials across all groups in which the algorithm's binary group decision is yes). Given that one-half of the trials were signal and that one-half were noise, the PC of each algorithm is equivalent to 50% of the algorithm's hit rate plus 50% of one minus the algorithm's false alarm rate. Each algorithm's hit rate and false alarm rate were used to estimate the algorithm's index of detectability as follows (60, 61): $d' = \Phi^{-1}(\text{hit rate}) - \Phi^{-1}(\text{false alarm rate})$, where Φ^{-1} is the inverse cumulative distribution function of the standard normal distribution.

Observers. Twenty undergraduates at the University of California, Santa Barbara volunteered or participated for course credit. All observers ran individually but saw the same images. Each observer ran in the search task first followed by the single-location task on a different day. Observers were 13 women and 7 men between 19 and 23 y old, except for one who was 28 y. One additional observer was omitted from the analyses, because his or her performance in the single-location task was close to chance. All observers were naïve to the purpose of the study. Procedures approved by the Human Subjects Committee at the University of California, Santa Barbara, were followed, and informed consent was obtained from all observers.

Eye Tracking. The gaze position of the observer's left eye was recorded at 250 Hz using a Desktop Mount EyeLink 1000 system (SR Research Ltd.). A calibration procedure was conducted at the start of each session using a nine-point grid system. The observer was allowed to recalibrate the eye tracker at any time. Saccades were classified as events in which eye velocity was higher than 35°/s and eye acceleration exceeded 9,500°/s^2.

Yes–No Signal Detection Task. The observer was informed that the target (a Gabor luminance patch) would be present in 50% of the trials. In the single-location task (Fig. 1*C*), the observer responded whether there was a target in the middle of the black square [known location paradigm (89)]. In the search task (Fig. 1*D*), the observer responded whether there was a target anywhere in the image [spatial uncertainty paradigm (89)]. The image for each trial was presented for 2,000 ms. The observer responded yes (i.e., target present) or no (i.e., target absent) by clicking on an eight-point confidence scale, where eight indicates a very high confidence that the target was present and one indicates a very high confidence that the target was absent. The eight confidence ratings appeared from left to right in color-coded boxes. The boxes for one through four were colored red to indicate no, and the boxes for five through eight were colored green to indicate yes.

Feedback. Response feedback was provided by lightly shading the four boxes that counted as correct responses for that trial (one through four for noise trials and five through eight for signal trials). In addition, if it was a signal trial, the stimulus was reshown for 1,500 ms with a small blue circle (0.672° diameter) around the target, so that the observer could learn and remember what the target looked like [signal known exactly paradigm (90, 91)].

Stimuli. The stimuli were created and displayed using MATLAB and the Psychophysics Toolbox libraries (92, 93). Each stimulus was an 8-bit gray-scale image that subtended ~27.72° × 22.33° and covered the entire 37.5- × 30-cm display (Barco MDRC-1119 LCD monitor; viewing distance = 76 cm; width × height resolution = 1,280 × 1,024 pixels; 1 pixel ~ 0.022°). The display was linearly calibrated to a mean luminance of 55.17 cd/m^2, with a minimum luminance of 0.07 cd/m^2 and a maximum luminance of 110.27 cd/m^2, and the ambient room lights were turned off.

Each image contained distinct additive luminance white noise (μ = 55.17 cd/m^2; σ = 8.61 cd/m^2; noise rms contrast = σ/μ = 0.156). The target, when present, was a vertically oriented sinusoidal luminance grating (spatial frequency = six cycles per 1°) enclosed in a Gaussian envelope (spatial SD = 0.12°). This luminance patch, known as a Gabor, was placed at a random location in the image. The center of the Gabor was restricted to be at least 4° away from the center of the image and at least 1.5° away from the edges of the image.

The energy of the signal (E) is defined as the sum of the squared luminance values of the entire Gabor as follows (94): $E = \sum\sum S(x,y)^2$, where $S(x,y)$ is the luminance value of the Gabor at each pixel location. The SNR is the distance in SD units between an ideal observer's respective decision variable distributions for target-present images and target-absent images. For white noise, it can be calculated from the signal and noise as follows (63, 95): SNR = root signal energy/noise SD = $\sqrt{E}/\sigma$.

The SD of the white noise was the same in both tasks (σ = 8.61 cd/m^2). The energy of the signal was manipulated between tasks by changing the contrast of the target. The peak contrast of the Gabor was 8% for the single-location task (E = 910.38 cd/m^2; SNR = 3.5) and 18% for the search task (E = 4,608.8 cd/m^2; SNR = 7.88).

In the search task, the location of the target was unknown to the observer. In the single-location task, the image contained a black square (1.5° × 1.5°) to indicate where the target might be as follows: during signal trials, the square was centered on the Gabor; during noise trials, the square was placed at a random location following the restrictions of where the center of the Gabor could appear during signal trials (at least 4° away from the center of the image and at least 1.5° away from the edges of the image).

Familiarization. All observers saw the same 100 example images in different random order. Unlike the practice and experimental images (see below), the target was present in 100% of the example images. The example images were distinct from the practice and experimental images but generated the same way. Observers' eyes were not tracked, and they did not make any responses. To illustrate where the target could appear (see above), the experimenter pointed to the edges and the center of the display during the first few examples and informed the observer that the target would never appear near the edges or the center of the image.

Each image was shown for 2,000 ms, after which a small blue circle (0.672° diameter) appeared to indicate where the target was located. The small blue circle was redundant with the black square in the single-location task (because the target was present in all of the example images), but we wanted to minimize the differences between the procedures of the two tasks. After the small blue circle appeared, the image remained on the display until the observer pressed a key to see the next example.

Practice Trials. All observers saw the same 50 practice images in different random order. The target was present in 50% of the practice images. The practice images were distinct from the experimental images but generated the same way. The practice trials were run in a single block before the experimental blocks, and observers were informed that the first block was practice. Observers' eyes were tracked, and they responded using the eight-point confidence scale. The experimenter explained the trial procedure (see below) during the first few practice trials.

Experimental Trials. All observers saw the same 500 experimental images in different random order. The target was present in 50% of the experimental images. Observers' eyes were tracked, and they responded using the eight-point confidence scale. The experimental images were divided into 10 blocks of 50 trials (for a total of 11 blocks including the practice block). Observers were encouraged to take short breaks between blocks.

Trial Procedure. The display was set to its mean luminance with a central fixation cross. Each trial began by pressing a key. After pressing the key, the observer had to maintain fixation within 1.1° of the fixation cross for a random delay of 500–1,000 ms before the onset of the image. This randomness was done to avoid anticipatory saccade planning to where the target could appear. If the observer did not maintain fixation on the central cross for the required amount of time, a message saying "broken fixation" would appear, and the observer had to press the key again to restart the trial.

After fixation was maintained for the required amount of time, the central cross would disappear to indicate that the image had appeared, and the observer was allowed to freely move his or her eyes. The image was shown for 2,000 ms before disappearing. The display was then reset to its mean luminance, and the eight-point confidence scale appeared in a random location. This randomness was done to avoid anticipatory eye movements to the confidence scale while the image was still shown. To respond, the observer clicked on his or her confidence rating for that trial and received immediate feedback as mentioned above. The observer cleared the response feedback by pressing a key. If it was a signal trial, the observer received additional feedback by reshowing the image with a circle around the target for 1,500 ms as mentioned above.

ACKNOWLEDGMENTS. We thank Craig Abbey for many helpful discussions. This study was supported by the Intelligence Community Postdoctoral Research Fellowship Program through Office of the Director of National Intelligence Grant 2011-11071400005, National Geospatial-Intelligence Agency Grant HM04761610003, and the Institute for Collaborative Biotechnologies through US Army Research Office Grant W911NF-09–0001. The content of the information does not necessarily reflect the position or the policy of the US Government, and no official endorsement should be inferred.

1. Beckers R, Deneubourg JL, Goss S (1992) Trails and U-turns in the selection of a path by the ant *Lasius niger*. *J Theor Biol* 159:397–415.
2. Seeley TD, Buhrman SC (1999) Group decision making in swarms of honey bees. *Behav Ecol Sociobiol* 45:19–31.
3. Amé J-M, Halloy J, Rivault C, Detrain C, Deneubourg JL (2006) Collegial decision making based on social amplification leads to optimal group formation. *Proc Natl Acad Sci USA* 103:5835–5840.
4. Sasaki T, Pratt SC (2012) Groups have a larger cognitive capacity than individuals. *Curr Biol* 22:R827–R829.
5. Sumpter DJT, Krause J, James R, Couzin ID, Ward AJW (2008) Consensus decision making by fish. *Curr Biol* 18:1773–1777.
6. Ward AJW, Sumpter DJT, Couzin ID, Hart PJB, Krause J (2008) Quorum decision-making facilitates information transfer in fish shoals. *Proc Natl Acad Sci USA* 105:6948–6953.
7. Ward AJW, Herbert-Read JE, Sumpter DJT, Krause J (2011) Fast and accurate decisions through collective vigilance in fish shoals. *Proc Natl Acad Sci USA* 108:2312–2315.
8. Black JM (1988) Preflight signalling in swans: A mechanism for group cohesion and flock formation. *Ethology* 79:143–157.
9. Simons AM (2004) Many wrongs: The advantage of group navigation. *Trends Ecol Evol* 19:453–455.
10. Dell'Ariccia G, Dell'Omo G, Wolfer DP, Lipp H-P (2008) Flock flying improves pigeons' homing: GPS track analysis of individual flyers versus small groups. *Anim Behav* 76: 1165–1172.
11. Prins HHT (1996) *Ecology and Behaviour of the African Buffalo* (Chapman & Hall, London).
12. Kerth G, Ebert C, Schmidtke C (2006) Group decision making in fission-fusion societies: Evidence from two-field experiments in Bechstein's bats. *Proc Biol Sci* 273:2785–2790.
13. Fischhoff IR, et al. (2007) Social relationships and reproductive state influence leadership roles in movements of plains zebra, *Equus burchellii*. *Anim Behav* 73:825–831.
14. Gautrais J, Michelena P, Sibbald A, Bon R, Deneubourg J-L (2007) Allelomimetic synchronization in Merino sheep. *Anim Behav* 74:1443–1454.
15. Stewart KJ, Harcourt AH (1994) Gorillas' vocalizations during rest periods: Signals of impending departure? *Behaviour* 130:29–40.
16. Leca J-B, Gunst N, Thierry B, Petit O (2003) Distributed leadership in semifree-ranging white-faced capuchin monkeys. *Anim Behav* 66:1045–1052.
17. Sueur C, Petit O (2008) Shared or unshared consensus decision in macaques? *Behav Processes* 78:84–92.
18. Strandburg-Peshkin A, Farine DR, Couzin ID, Crofoot MC (2015) Shared decision-making drives collective movement in wild baboons. *Science* 348:1358–1361.
19. Conradt L, Roper TJ (2003) Group decision-making in animals. *Nature* 421:155–158.
20. Conradt L, Roper TJ (2005) Consensus decision making in animals. *Trends Ecol Evol* 20:449–456.
21. Schaller GB (1963) *The Mountain Gorilla: Ecology and Behavior* (Univ of Chicago Press, Chicago).
22. Dumont B, Boissy A, Achard C, Sibbald AM, Erhard HW (2005) Consistency of animal order in spontaneous group movements allows the measurement of leadership in a group of grazing heifers. *Appl Anim Behav Sci* 95:55–66.

23. King AJ, Douglas CMS, Huchard E, Isaac NJB, Cowlishaw G (2008) Dominance and affiliation mediate despotism in a social primate. *Curr Biol* 18:1833–1838.
24. Sorkin RD, West R, Robinson DE (1998) Group performance depends on the majority rule. *Psychol Sci* 9:456–463.
25. Sorkin RD, Hays CJ, West R (2001) Signal-detection analysis of group decision making. *Psychol Rev* 108:183–203.
26. Couzin ID, Krause J, Franks NR, Levin SA (2005) Effective leadership and decision-making in animal groups on the move. *Nature* 433:513–516.
27. Kalven H, Zeisel H (1966) *The American Jury* (Little, Brown and Co., Boston).
28. Boehm C (1996) Emergency decisions, cultural-selection mechanics, and group selection. *Curr Anthropol* 37:763–793.
29. Millstein IM, MacAvoy PW (1998) The active board of directors and performance of the large publicly traded corporation. *Columbia Law Rev* 98:1283–1322.
30. Riemer N (1951) The case for bare majority rule. *Ethics* 62:16–32.
31. Hastie R, Kameda T (2005) The robust beauty of majority rules in group decisions. *Psychol Rev* 112:494–508.
32. Condorcet M (1785) *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix* (Imprimerie Royale, Paris).
33. Sorkin RD, Luan S, Itzkowitz J (2008) Group decision and deliberation: A distributed detection process. *Blackwell Handbook of Judgment and Decision Making*, eds Koehler DJ, Harvey N (Wiley, New York), pp 464–484.
34. Juni MZ, Eckstein MP (2015) Flexible human collective wisdom. *J Exp Psychol Hum Percept Perform* 41:1588–1611.
35. Waldron J (2014) Five to four: Why do bare majorities rule on courts. *Yale Law J* 123: 1692–1730.
36. Galton F (1907) Vox populi. *Nature* 75:450–451.
37. Gordon K (1924) Group judgments in the field of lifted weights. *J Exp Psychol* 7: 398–400.
38. Bruce RS (1935) Group judgments in the fields of lifted weights and visual discrimination. *J Psychol* 1:117–121.
39. Smith M, Wilson EA (1953) A model of the auditory threshold and its application to the problem of the multiple observer. *Psychol Monogr Gen Appl* 67(9):1–35.
40. Clement DE, Schiereck JJ (1973) Sex composition and group performance in a visual signal detection task. *Mem Cognit* 1:251–255.
41. Metz CE, Shen J-H (1992) Gains in accuracy from replicated readings of diagnostic images: Prediction and assessment in terms of ROC analysis. *Med Decis Making* 12: 60–75.
42. Jiang Y, Nishikawa RM, Schmidt RA, Metz CE (2006) Comparison of independent double readings and computer-aided diagnosis (CAD) for the diagnosis of breast calcifications. *Acad Radiol* 13:84–94.
43. Bahrami B, et al. (2010) Optimally interacting minds. *Science* 329:1081–1085.
44. Kurvers RHJM, et al. (2016) Boosting medical diagnostics by pooling independent judgments. *Proc Natl Acad Sci USA* 113:8777–8782.
45. Smith M (1931) Group judgments in the field of personality traits. *J Exp Psychol* 14: 562–565.
46. Libby R, Blashfield RK (1978) Performance of a composite as a function of the number of judges. *Organ Behav Hum Perform* 21:121–129.
47. Eckstein MP, et al. (2012) Neural decoding of collective wisdom with multi-brain computing. *Neuroimage* 59:94–108.
48. Bates JM, Granger CWJ (1969) The combination of forecasts. *Oper Res Q* 20: 451–468.
49. Clemen RT (1989) Combining forecasts: A review and annotated bibliography. *Int J Forecast* 5:559–583.
50. Mozer MC, Pashler H, Homaei H (2008) Optimal predictions in everyday cognition: The wisdom of individuals or crowds? *Cogn Sci* 32:1133–1147.
51. Hueffer K, Fonseca MA, Leiserowitz A, Taylor KM (2013) The wisdom of crowds: Predicting a weather and climate-related event. *Judgm Decis Mak* 8:91–105.
52. Kattan MW, O'Rourke C, Yu C, Chagin K (2016) The wisdom of crowds of doctors: Their average predictions outperform their individual ones. *Med Decis Making* 36: 536–540.
53. Surowiecki J (2005) *The Wisdom of Crowds* (Anchor Books, New York).
54. Green DM, Swets JA (1966) *Signal Detection Theory and Psychophysics* (Wiley, New York).
55. Lee EH, Jun JK, Jung SE, Kim YM, Choi N (2014) The efficacy of mammography boot camp to improve the performance of radiologists. *Korean J Radiol* 15:578–585.
56. Krupinski EA (2010) Current perspectives in medical image perception. *Atten Percept Psychophys* 72:1205–1217.
57. Schafer TH (1949) *Detection of a Signal by Several Observers* (US Naval Electronics Laboratory, San Diego), USNEL Rep No 101.
58. Hogarth RM (1978) A note on aggregating opinions. *Organ Behav Hum Perform* 21: 40–46.
59. Sorkin RD, Dai H (1994) Signal detection analysis of the ideal group. *Organ Behav Hum Decis Process* 60:1–13.
60. Abbey CK, Bochud FO (2000) Modeling visual detection tasks in correlated image noise with linear model observers. *Handbook of Medical Imaging, Volume 1. Physics and Psychophysics*, eds Beutel J, Kundel HL, Van Metter RL (SPIE Press, Bellingham, WA), pp 629–654.
61. Wickens TD (2001) *Elementary Signal Detection Theory* (Oxford Univ Press, New York).
62. Tanner WP, Jr, Birdsall TG (1958) Definitions of d' and η as psychophysical measures. *J Acoust Soc Am* 30:922–928.
63. Burgess AE, Wagner RF, Jennings RJ, Barlow HB (1981) Efficiency of human visual signal discrimination. *Science* 214:93–94.
64. Barr S, Gold JM (2014) Redundant visual information enhances group decisions. *J Exp Psychol Hum Percept Perform* 40:2124–2130.
65. Eckstein MP, Beutter BR, Stone LS (2001) Quantifying the performance limits of human saccadic targeting during visual search. *Perception* 30:1389–1401.
66. Peterson MF, et al. (2010) Ideal observer analysis for task normalization of pattern classifier performance applied to EEG and fMRI data. *J Opt Soc Am A Opt Image Sci Vis* 27:2670–2683.
67. Efron B, Tibshirani R (1993) *An Introduction to the Bootstrap* (Chapman & Hall, New York).
68. Peli E, Yang J, Goldstein RB (1991) Image invariance with changes in size: The role of peripheral contrast thresholds. *J Opt Soc Am A* 8:1762–1774.
69. Macke JH, Berens P, Ecker AS, Tolias AS, Bethge M (2009) Generating spike trains with specified correlation coefficients. *Neural Comput* 21:397–423.
70. Malcolmson KA, Reynolds MG, Smilek D (2007) Collaboration during visual search. *Psychon Bull Rev* 14:704–709.
71. Brennan SE, Chen X, Dickinson CA, Neider MB, Zelinsky GJ (2008) Coordinating cognition: The costs and benefits of shared gaze during collaborative search. *Cognition* 106:1465–1477.
72. Brennan AA, Enns JT (2015) When two heads are better than one: Interactive versus independent benefits of collaborative cognition. *Psychon Bull Rev* 22:1076–1082.
73. Brady CJ, et al. (2014) Rapid grading of fundus photographs for diabetic retinopathy using crowdsourcing. *J Med Internet Res* 16:e233.
74. Mitry D, et al. (2015) Crowdsourcing as a screening tool to detect clinical features of glaucomatous optic neuropathy from digital photography. *PLoS One* 10:e0117401.
75. Nguyen TB, et al. (2012) Distributed human intelligence for colonic polyp classification in computer-aided detection for CT colonography. *Radiology* 262:824–833.
76. Irshad H, et al. (2015) Crowdsourcing image annotation for nucleus detection and segmentation in computational pathology: Evaluating experts, automated methods, and the crowd. *Pac Symp Biocomput* 2015:294–305.
77. Liu B, Metz CE, Jiang Y (2004) An ROC comparison of four methods of combining information from multiple images of the same patient. *Med Phys* 31:2552–2563.
78. Bradley C, Abrams J, Geisler WS (2014) Retina-V1 model of detectability across the visual field. *J Vis* 14:22.
79. Drew T, et al. (2013) Scanners and drillers: Characterizing expert visual search through volumetric images. *J Vis* 13:3.
80. Rubin GD, et al. (2015) Characterizing search, recognition, and decision in the detection of lung nodules on CT scans: Elucidation with eye tracking. *Radiology* 274: 276–286.
81. Deza A, Eckstein M (2016) Can peripheral representations improve clutter metrics on complex scenes? *Advances in Neural Information Processing Systems*, eds Lee DD, Luxburg UV, Guyon I, Garnett R (Curran Associates, Inc., Red Hook, NY), Vol 29, pp 2847–2855.
82. Diaz I, Eckstein MP, Luyet A, Bize P, Bochud FO (2012) Measurements of the detectability of hepatic hypovascular metastases as a function of retinal eccentricity in CT images. *Proc SPIE* 8318:0J-1–0J-6.
83. Rogalla P, Kloeters C, Hein PA (2009) CT technology overview: 64-Slice and beyond. *Radiol Clin North Am* 47:1–11.
84. Good WF, et al. (2008) Digital breast tomosynthesis: A pilot observer study. *AJR Am J Roentgenol* 190:865–869.
85. Gur D, et al. (2009) Digital breast tomosynthesis: Observer performance study. *AJR Am J Roentgenol* 193:586–591.
86. Baker JA, Lo JY (2011) Breast tomosynthesis: State-of-the-art and review of the literature. *Acad Radiol* 18:1298–1310.
87. Soll JB, Larrick RP (2009) Strategies for revising judgment: How (and how well) people use others' opinions. *J Exp Psychol Learn Mem Cogn* 35:780–805.
88. Prelec D, Seung HS, McCoy J (2017) A solution to the single-question crowd wisdom problem. *Nature* 541:532–535.
89. Swensson RG, Judy PF (1981) Detection of noisy visual targets: Models for the effects of spatial uncertainty and signal-to-noise ratio. *Percept Psychophys* 29:521–534.
90. Peterson WW, Birdsall TG, Fox WC (1954) The theory of signal detectability. *Trans IRE Prof Group Inf Theory* 4(4):171–212.
91. Tanner WP, Jr, Swets JA (1954) The human use of information–I: Signal detection for the case of the signal known exactly. *Trans IRE Prof Group Inf Theory* 4(4):213–221.
92. Brainard DH (1997) The psychophysics toolbox. *Spat Vis* 10:433–436.
93. Pelli DG (1997) The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spat Vis* 10:437–442.
94. Burgess AE, Colborne B (1988) Visual signal detection. IV. Observer inconsistency. *J Opt Soc Am A* 5:617–627.
95. Watson AB, Barlow HB, Robson JG (1983) What does the eye see best? *Nature* 302: 419–422.
96. Fukunaga K (1990) *Introduction to Statistical Pattern Recognition* (Academic, London), 2nd Ed.