Psychon Bull Rev
DOI 10.3758/s13423-013-0459-3

BRIEF REPORT

The gist of the abnormal: Above-chance medical decision making in the blink of an eye

Karla K. Evans · Diane Georgian-Smith · Rosemary Tambouret · Robyn L. Birdwell · Jeremy M. Wolfe

Abstract Very fast extraction of global structural and statistical regularities allows us to access the "gist"—the basic meaning—of real-world images in as little as 20 ms. Gist processing is central to efficient assessment and orienting in complex environments. This ability is probably based on our extensive experience with the regularities of the natural world. If that is so, would experts develop an ability to extract the gist from the artificial stimuli (e.g., medical images) with which they have extensive visual experience? Anecdotally, experts report some ability to categorize images as normal or abnormal before actually finding an abnormality. We tested the reality of this perception in two expert populations: radiologists and cytologists. Observers viewed brief (250- to 2,000-ms) presentations of medical images. The presence of abnormality was randomized across trials. The task was to rate the abnormality of an image on a 0–100 analog scale and then to attempt to localize that abnormality on a subsequent screen showing only the outline of the image. Both groups of experts had above-chance performance for detecting subtle abnormalities at all stimulus durations (cytologists $d' \approx 1.2$ and radiologists $d' \approx 1$), whereas the nonexpert control groups did not differ from chance ($d' \approx 0.23$, $d' \approx 0.25$). Furthermore, the experts' ability to localize these abnormalities was at chance levels, suggesting that categorization was based on a global signal, and not on fortuitous attention to a localized target. It is possible that this global signal could be exploited to improve clinical performance.

Keywords Visual awareness · Human visual perception · Visual categorization · Expertise

Our visual world is very rich and complex, providing us with more information than our visual system can handle. Nevertheless, in spite of the limitations of visual processing, we are still able to perceive significant information about a scene after a fraction of a second's exposure to it. An exposure on the order of 100 ms enables us to assess the general meaning or "gist" of a completely novel scene (Intraub, 1981; Potter & Faulconer, 1975). Likewise, a 20-ms masked exposure is enough to categorize the basic (e.g., lake vs. forest) or superordinate (e.g., natural vs. urban) level of a scene with above-chance accuracy (Greene & Oliva, 2009; Joubert, Rousselet, Fize, & Fabre-Thorpe, 2007). If primed with a category (e.g., animal), observers are above chance at the detection of large objects (Thorpe, Fize, & Marlot, 1996; VanRullen & Thorpe, 2001), even when focused attention is occupied with another foveal task (Li, VanRullen, Koch, & Perona, 2002). In fact, observers are capable of rapidly extracting information about *multiple* categories, even if they do not know the target category (animal, beach, mountain, etc.) in advance (Evans, Horowitz, & Wolfe, 2011). These abilities appear to be based on the interpretation of global properties and image statistics from our experience with the regularities in the natural world (Evans et al., 2011; Wolfe, Võ, Evans, & Greene, 2011).

Medical experts, performing complex perceptual tasks with man-made visual stimuli, sometimes report that they feel as if they can categorize an image as normal or abnormal in a single glance. In the medical image perception literature are reports

K. K. Evans · D. Georgian-Smith · R. L. Birdwell · J. M. Wolfe
Brigham and Women's Hospital, Boston, MA, USA

K. K. Evans · D. Georgian-Smith · R. Tambouret ·
R. L. Birdwell · J. M. Wolfe
Harvard Medical School, Boston, MA, USA

R. Tambouret
Massachusetts General Hospital, Boston, MA, USA

K. K. Evans (✉)
Visual Attention Lab, Harvard Medical School,
64 Sidney Street, Suite 170,
Cambridge, MA 02139, USA
e-mail: kevans@search.bwh.harvard.edu

Published online: 15 June 2013

of radiologists detecting lesions in chest radiographs and mammograms at above-chance levels with only a quarter of a second's glimpse of the image (Carmody, Nodine, & Kundel, 1981; Kundel & Nodine, 1975; Mugglestone, Gale, Cowley, & Wilson, 1995; Oestmann et al., 1988). Kundel and his colleagues have been leaders in the study of the first moments of medical image perception: What does an expert get from the first glimpse of an image? They have developed "the hypothesis that visual search begins with a global response that establishes content, detects gross deviations from normal, and organizes subsequent foveal checking fixations" (Kundel & Nodine, 1975, p. 527). In this article, we will focus on what might be considered the aspect of early processing that "establishes content." We hypothesized that experts can also sense a global signal, akin to the signals that allow for rapid natural-scene categorization. This signal would not necessarily "organize subsequent fixations" and localize suspicious regions, as hypothesized by Kundel and Nodine. Rather, it might contribute to a conviction that a subsequent search would uncover an abnormality, thus changing the searcher's approach to the task. We suggest that this sort of global, gist signal would be based on an implicit extraction of statistics across the whole image. As we will demonstrate, such a signal could allow for above-chance categorization of the image without supporting effective object recognition within the image, nor would it necessarily constrain future eye movements, although in the present work, we did not look at eye movements.

To evaluate this hypothesis, we tested two sets of medical experts on their ability to extract the gist of the "abnormal" using briefly presented images from their domain of expertise. A total of 55 radiologists were presented with 100 trials of craniocaudal or mediolateral oblique X-ray views of both breasts (Fig. 1a), and 38 cytologists saw 120 Pap test images (micrographs of many cervical cells; Fig. 1b). The exposures were from 250 to 2,000 ms in duration, with presence of abnormality and duration randomized across trials. Observers rated the abnormality of an image on a 0–100 analog scale. Half of the cases were verified as not having any abnormality, and the other half had various subtle abnormalities. For each expert group, we tested a control group of naïve observers who had no significant experience with images of these sorts.

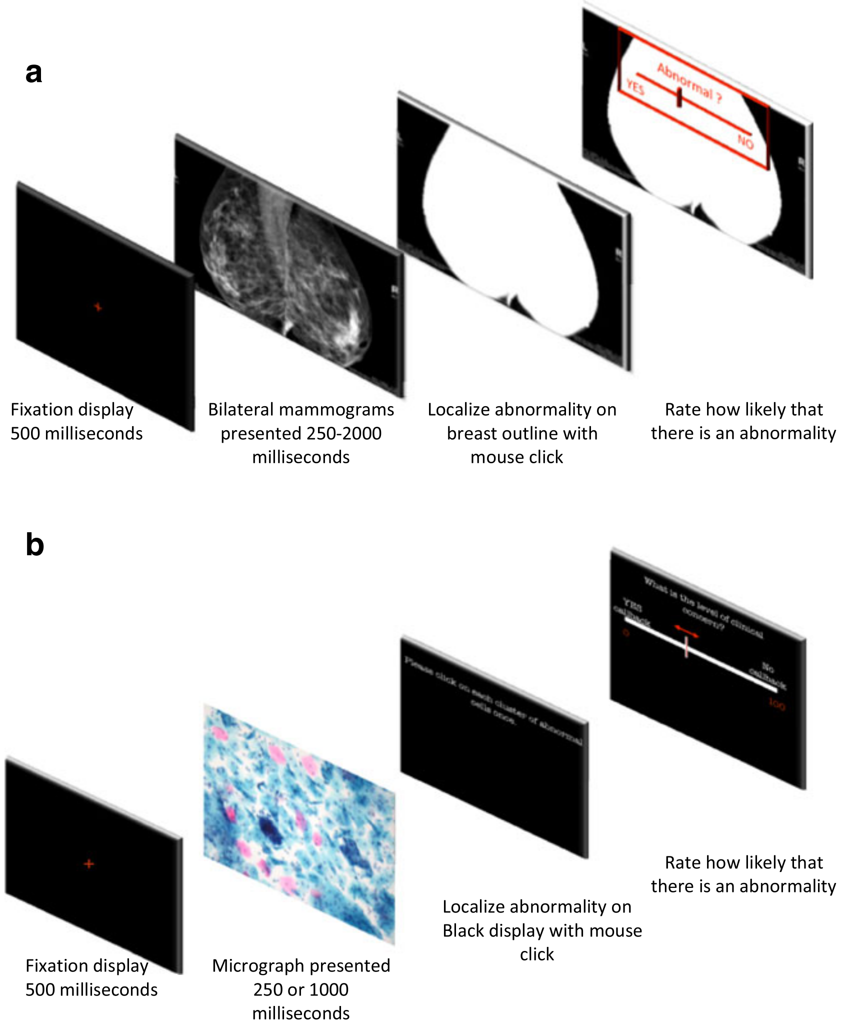

Fig. 1 Example trials in our three experiments. (a) Example trial presented to radiologists and a naïve control group in Experiments 1 and 2. (b) Example trial presented to cytologists and a naïve control group in Experiment 3

Materials and method

Participants

We tested 55 radiologists (32 female, 23 male; average age 56 years), 38 cytologists (22 female, 16 male; average age 51 years), and 60 nonexpert control observers (36 female, 24 male; average age 30 years). All had normal or corrected-to-normal vision and gave informed consent. All medical expert recruits were actively engaged in the daily practice of laboratory cervical cytology and radiology screening and had at least 5 years of experience. The group of radiologist experts had an average of 18 years of experience and an estimated range of 1,000–15,000 cases diagnosed per year. The group of cytologists had an average of 21 years of experience and an estimated range of 1,500–18,000 cases diagnosed per year. The naïve observers were recruited from the greater Boston area, had no medical training, and were randomly assigned to view either cytology or mammography images. They performed the same task as the experts after a short tutorial.

Stimuli and procedure

All observers viewed the images for a very brief time, with durations ranging from 250 to 2,000 ms. Most of the observers saw images in a mixture of two durations. In the first experiment, 20 radiologists saw 100 images for 500 and 1,000 ms, 13 for 250 and 2,000 ms, and seven for 750 ms. In Experiment 2, 15 radiologists saw the 100 images, all of them for 500 ms, and the control group of 30 naïve observers saw the same images for 250 and 1,000 ms. In Experiment 3, the entire expert group of cytologists and the control group of 30 naïve observers saw images for 250 and 1,000 ms. In Experiments 1 and 2, after ten practice trials, all of the expert groups of radiologists and their control group completed 100 trials in which they viewed craniocaudal or mediolateral oblique views of X-ray images of breasts (mammograms). The expert group of cytologists and their control group completed 120 trials viewing Pap test images (micrographs) of many cervical cells. All three experiments were conducted on a Dell Studio computer, and the images were displayed on 19-in., liquid-crystal color display screen at a viewing distance of 53 cm. The resolution of the monitor was 1,920 × 1,200 pixels, with a usable intensity range of 2–260 candelas per square meter and a contrast ratio of 188:1. Half of the images were normal, and half showed cancerous abnormalities. In mammograms, the cases considered as having abnormalities were mammograms that were either screen-detected cancers, confirmed by later biopsy, or mammograms that had been done 1 to 2 years prior to a screen-detected cancer and had been interpreted as negative. In the latter cases, the lesion was retroactively determined by a study radiologist to have been visible. The abnormalities in the mammograms were subtle masses and architectural distortions, and the subtlety of the findings was determined by the study radiologists on the basis of their long years of experience in the practice. The average size of the lesions in the mammograms was 18 millimeters, and the range was from 10 to 48 mm. The abnormalities in the micrographs were both low-grade and high-grade squamous intraepithelial lesions. As was the case with the mammograms, 50 % of the micrographs had abnormalities, and on average 19 % of the slide showed abnormality.

The mammography images were full-field digital mammograms, and the cytology images were digitized glass-slide samples that were examined and photographed at 3,100 magnification (310 ocular and 310 objective) using an Olympus BX51 optical microscope (Olympus, Tokyo, Japan) equipped with a ColorView II digital camera (Soft Imaging System Ltd, Helperby, North Yorkshire, England). The starting resolution of the mammograms was 1,980 × 2,294 pixels, but they were then downsized to 900 × 1,150 pixels so that two could be presented at a time, side by side. The micrographs were acquired using analysis software (Soft Imaging System) at a resolution of 2,010 × 51,544 pixels, but they were then downsized to a resolution of 2,080 × 1,544 pixels for presentation purposes. In addition, after viewing, but before rating the images, the 15 radiologists in the second experiment and all of the cytologists and their control group were asked to localize the abnormalities by clicking on the display screen where they thought that they saw an abnormality. The micrographs were followed by a blank, black screen, and the mammograms by an outline of the breasts on which the observers were asked to make their best guess as to the location of the abnormalities.

Data analysis

We measured performance in terms of d' values derived from the confidence ratings, which ranged from 0 to 100. We used the rating of 50 as a neutral divider of the ratings to convert them into binary responses of “YES, there is an abnormality” or “NO, there is no abnormality.” We adopted this measure for two reasons. First, d' is theoretically independent of an observer’s bias to respond “yes” or “no.” Second, it is normally distributed, unlike accuracy, which makes it more suitable for standard parametric statistics. In addition to d', we calculated a related measure derived from the ratings: the area under curve (AUC) of the receiver operating characteristic curve (ROC).

Localization performance was measured by determining the percentage of observers’ clicks inside predetermined regions of interest (ROIs) delineating the abnormalities. The ROI’s were drawn around all the abnormalities whether they were in a cluster or individual cells for micrographs and around the entire lesion (mass or architectural distortion) in mammograms. Localization was assessed for trials on which observers correctly rated the images as abnormal. Chance levels for localization performance were determined by calculating the average percentage of overall tissue area covered by an ROI.

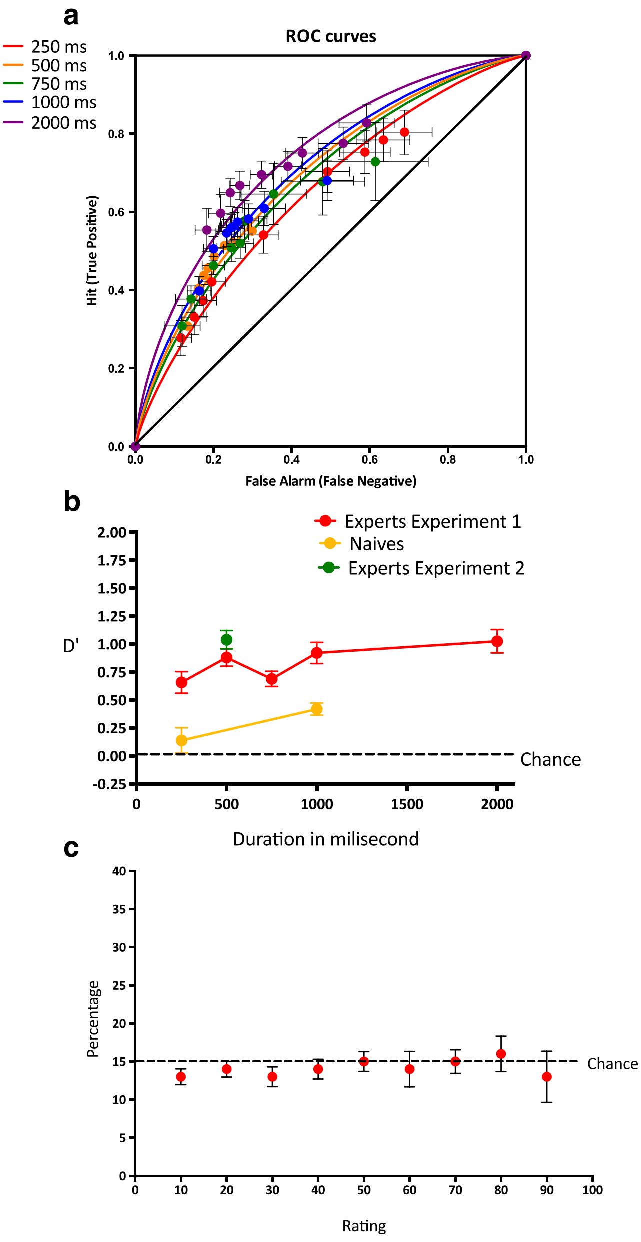

Fig. 2 Results of Experiments 1 and 2. (a) Rating-based receiver operating characteristic (ROC) curves of radiologists’ performance for each of the five exposure durations in Experiment 1. (b) Performance of the expert groups (radiologists in Exps. 1 and 2) and the control group (naïve observers), measured in *d′* units for each exposure duration. (c) Localization performance of radiologists in Experiment 2 across different confidence ratings. The error bars in panels a, b and c are standard errors of the means

Results and discussion

Both groups of experts had above-chance performance for detecting subtle abnormalities at all stimulus durations (cytologists $d' \approx 1.2$ and radiologists $d' \approx 1$; Figs. 2a and 3a). Neither of the control groups achieved performance significantly better than chance at the short duration of 250 ms, and they were considerably worse than the experts

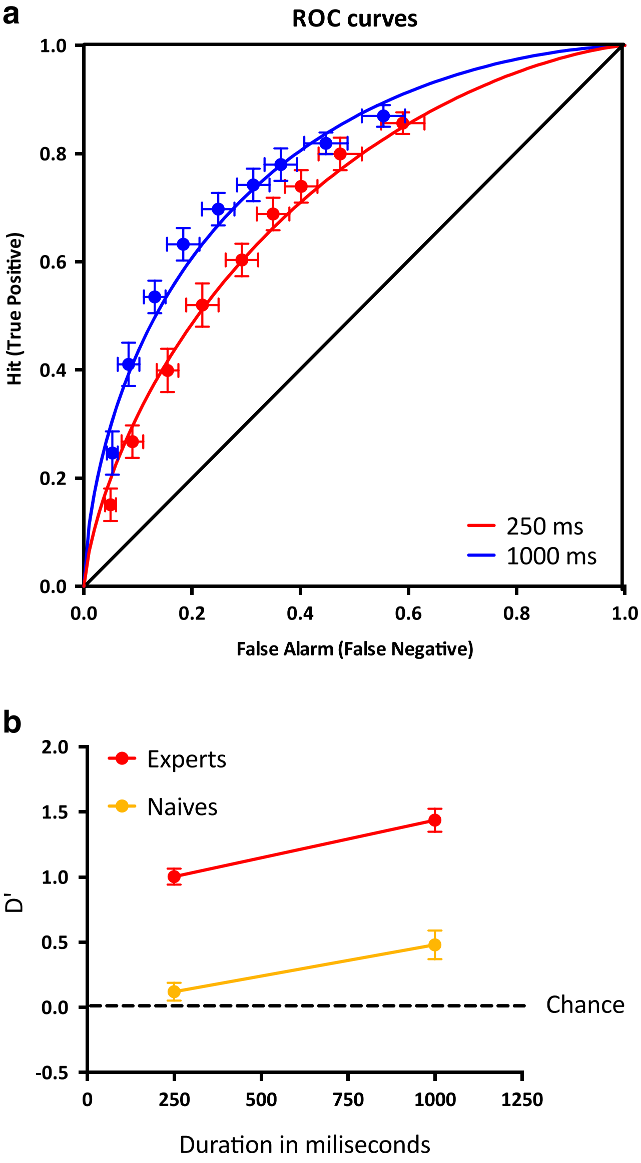

Fig. 3 Results of Experiment 3. (a) ROCs showing cytologists’ performance for each of the two exposure durations. (b) Performance of the expert group (cytologists) and control group (naïve observers), measured in *d′* units for the two exposure durations. The error bars in panels a and b are standard errors of the means

at 1,000 ms (Figs. 2b and 3b). For the radiologists, tested at five different exposure durations, *t* tests on *d′* showed that performance was significantly above chance at each exposure duration [250 ms, *t*(19) = 6.82, *p* < .0001, AUC = 0.64; 500 ms, *t*(19) = 11.28, *p* < .0001, AUC = 0.65; 750 ms, *t*(7) = 10.11, *p* < .0001, AUC = 0.65; 1,000 ms, *t*(19) = 9.79, *p* < .0001, AUC = 0.66; 2,000 ms, *t*(12) = 9.86, *p* < .0001, AUC = 0.72] (see Fig. 2). Eye movements (e.g., to each breast) were not required for this rapid gist extraction, because the 250-ms condition did not permit volitional eye movements to each breast, but was not significantly worse than the 1,000-ms exposure [*t*(19) = 1.8137, *p* = .0856]. Cytologists showed a similar pattern of results, with above-chance performance at both exposure durations [250 ms, *t*(37) = 16.22, *p* < .0001, AUC = 0.71, and 1,000 ms, *t*(37) = 16.37, *p* < .0001, AUC = 0.77] (Fig. 3a and b). In this case, even the experts showed significant improvement at the longer duration [*t*(37) = 4.42, *p* < .0001].

Before rating each briefly exposed stimulus, all of the cytology experts and the smaller group of 15 radiologists in the second experiment were asked to localize abnormalities on a screen showing only the outline of the image (Fig. 4a). For a localization to be deemed "correct," it needed to fall within an ROI delineating the abnormality. These ROIs were defined for each image by one of the authors: D.G.-S., for mammography, and R.T., for cytology. We calculated the percentages of correctly localized abnormalities with respect to the overall numbers of abnormalities. Chance level was defined as the percentage of the image lying within the abnormal region—that is, the percentage that might be achieved by making random localizations. Localization performance for both groups was very poor, not significantly different from chance (cytologists showed 16 % correct localizations, and radiologists, 15 %; see Figs. 2c and 4a). Interestingly, the localization performance did not improve as the confidence rating increased, staying flat for both expert groups, and for both exposure durations in the cytology group.

This result should not be mistaken for a claim that we can or should make important decisions in the blink of an eye. No one would suggest performing cancer screening in 250 ms with a *d′* of ~1.0, given that the performance of expert radiologists is *d′* = 2.5–3.0 (e.g. D'Orsi et al., 2013). This result does not show that assessment is over in a blink of an eye. Rather, it shows that, with the correct training, experts can have a global impression of the normality or abnormality of a medical image. That impression appears to be based on a global signal that, by itself, is not sufficient to localize the target. Mack and Palmeri (2010) drew similar conclusions about real-scene categorization. They found that a computational model based solely on global scene statistics can explain the consistent-object advantage in rapid scene categorization without any explicit rapid object recognition within a scene. Thus, global image statistics are sufficient to differentiate scenes with objects that are consistent or inconsistent within the overall scene context. We believe that this is a trained specialization of normal gist processing. Experience with the world has taught all of us that *this* set of image statistics is typically associated with, for example, an urban street, whereas *that* set of image statistics is associated with farmland. With specific training, an expert radiologist or cytologist learns the statistical regularities that distinguish normal from abnormal in the images in their realm of expertise. The ability to have a feeling that something is amiss, and yet not know where to find it, is akin to what Rensink (2004) called "mindsight," in which observers consciously sense that a change has occurred but still have no visual experience of that change. However, the method of establishing mindsight, by looking at the time difference between when observers report a "sense" versus a "saw"

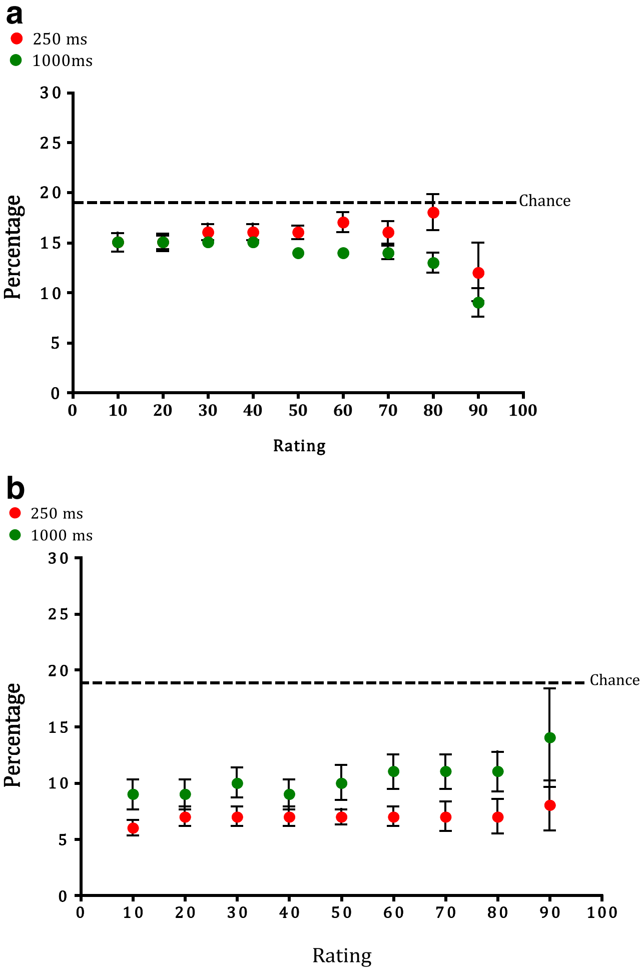

Fig. 4 Localization performance for (a) cytologists and (b) naïves in Experiment 3 across different confidence ratings and the two exposure durations. The error bars in panels a and b are standard errors of the means

response in a change detection task, has been put into question (Simons, Nevarez, & Boot, 2005). Simons et al. argued that the findings that Rensink attributed to mindsight could be explained as the results of a verification process, rather like a shift from an initial liberal to a more conservative decision criterion.

In the present study, a criterion shift alone would not account for the ability to perform at above-chance levels. Nevertheless, the utility of this interesting talent should not be oversold. For instance, it would be unwise in clinical practice to declare an image "'normal'" on the basis of the global signal. However, devoting extra scrutiny to images that *feel* abnormal might improve performance. In addition, more work must be done to relate these findings with earlier work on the ability of experts to use the first second or so of exposure to direct their eyes intelligently in subsequent search, as was proposed by Kundel's group . We did not measure eye movements, and indeed, at our shorter exposure durations, there would not have been much point in doing so. Had the stimulus remained visible, it is possible that the experts' eyes would have been guided, even though they were unable to use the initial information to localize targets. The relationship between this unlocalizable initial signal and the course of subsequent search remains to be worked out.

Finally, if the gist signal could be identified by a computer, it might be used as a novel form of computer-aided detection (CAD), even if it did not direct the gaze of the expert. Normal mammography CAD markings are used to indicate the possible locations of targets. A global CAD would simply be a warning that a particular case had a statistical signal that was associated with an elevated chance of a positive finding (see Hope et al., 2013). This signal would be similar to a genetic risk factor, telling the clinician that something was more likely to be present.

Author note This research was funded by a Ruth L. Kirschstein National Research Service Award, Grant No. F32EY019819-01 to K.K.E., and by National Institutes of Health–National Eye Institute Grant No. EY17001 and a grant from Toshiba to J.M.W.

References

Carmody, D. P., Nodine, C. F., & Kundel, H. L. (1981). Finding lung nodules with and without comparative visual scanning. *Perception & Psychophysics, 29,* 594–598.

D'Orsi, C. J., Getty, D. J., Pickett, R. M., Sechopoulos, I., Newell, M. S., Gundry, K. R., ... D'Orsi, E. M. (2013). Stereoscopic digital mammography: Improved specificity and reduced rate of recall in a prospective clinical trial. *Radiology, 266*(1), 81–88.

Evans, K. K., Horowitz, T. S., & Wolfe, J. M. (2011). When categories collide: Accumulation of information about multiple categories in rapid scene perception. *Psychological Science, 22,* 739–746. doi:10.1177/0956797611407930

Greene, M. R., & Oliva, A. (2009). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive Psychology, 58,* 137–176. doi:10.1016/j.cogpsych.2008.06.001

Hope, C. I., Sterr, A., Elangovan, P., Geades, N., Windridge, D., Young, K. C., & Wells, K. (2013). High throughput screening for mammography using a human–computer interface with rapid serial visual presentation (RSVP). In C. K. Abbey & C. R. Mello-Thoms (Eds.), *Medical Imaging 2013: Image perception, observer performance, and technology assessment (Proceedings of SPIE, Vol. 8673)*. doi:10.1117/12.2007557

Intraub, H. (1981). Rapid conceptual identification of sequentially presented pictures. *Journal of Experimental Psychology. Human Perception and Performance, 7,* 604–610. doi:10.1037/0096-1523.7.3.604

Joubert, O. R., Rousselet, G. A., Fize, D., & Fabre-Thorpe, M. (2007). Processing scene context: Fast categorization and object interference. *Vision Research, 47,* 3286–3297.

Li, F. F., VanRullen, R., Koch, C., & Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences, 99,* 9596–9601.

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

Mack, M. L., & Palmeri, T. J. (2010). Modeling categorization of scenes containing consistent versus inconsistent objects. *Journal of Vision, 10*(3), 11, 1–11. doi:10.1167/10.3.11

Mugglestone, M. D., Gale, A. G., Cowley, H. C., & Wilson, A. R. M. (1995). Diagnostic performance on briefly presented mammographic images. In H. L. Kundel (Ed.), *Medical Imaging 1995: Image perception (Proceedings of SPIE, Vol. 2436)* (pp. 106–116). Bellingham, WA: SPIE. doi:10.1117/12.206840

Oestmann, J. W., Greene, R., Kushner, D. C., Bourgouin, P. M., Linetsky, L., & Llewellyn, H. J. (1988). Lung lesions: Correlation between viewing time and detection. *Radiology, 166,* 451–453.

Potter, M. C., & Faulconer, B. A. (1975). Time to understand pictures and words. *Nature, 253,* 437–438.

Rensink, R. A. (2004). Visual sensing without seeing. *Psychological Science, 15,* 27–32.

Simons, D. J., Nevarez, G., & Boot, W. R. (2005). Visual sensing is seeing why "mindsight", in hindsight, is blind. *Psychological Science, 16,* 520–524.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381,* 520–522.

VanRullen, R., & Thorpe, S. J. (2001). Is it a bird? Is it a plane? Ultra-rapid visual categorisation of natural and artifactual objects. *Perception, 30,* 655–668.

Wolfe, J. M., Võ, M. L.-H., Evans, K. K., & Greene, M. R. (2011). Visual search in scenes involves selective and nonselective pathways. *Trends in Cognitive Sciences, 15,* 77–84. doi:10.1016/j.tics.2010.12.001