Journal of Experimental Psychology:
Human Perception and Performance
2015, Vol. 41, No. 1, 1–5

0096-1523/15/$12.00 http://dx.doi.org/10.1037/xhp0000012

OBSERVATION

What Can 1 Billion Trials Tell Us About Visual Search?

Stephen R. Mitroff, Adam T. Biggs, Stephen H. Adamo,
Emma Wu Dowd, Jonathan Winkle, and Kait Clark
Duke University

Mobile technology (e.g., smartphones and tablets) has provided psychologists with a wonderful opportunity: through careful design and implementation, mobile applications can be used to crowd source data collection. By garnering massive amounts of data from a wide variety of individuals, it is possible to explore psychological questions that have, to date, been out of reach. Here we discuss 2 examples of how data from the mobile game *Airport Scanner* (Kedlin Co., http://www.airportscannergame.com) can be used to address questions about the nature of visual search that pose intractable problems for laboratory-based research. *Airport Scanner* is a successful mobile game with millions of unique users and billions of individual trials, which allows for examining nuanced visual search questions. The goals of the current Observation Report were to highlight the growing opportunity that mobile technology affords psychological research and to provide an example roadmap of how to successfully collect usable data.

Keywords: visual search, big data, mobile applications, airport scanner, multiple-target search

Psychological researchers have never lacked for good questions. However, sometimes method and technology lag behind—thereby withholding the necessary tools for addressing certain questions. As such, each new technological and/or methodological advance provides the field with a possible means to move forward. Cognitive psychology, for example, has grown hand-in-hand with each new available technology (e.g., Wilhelm Wundt's laboratory devices, the tachistoscope, personal computers, eye-tracking devices, brain-imaging techniques).

At present, we are in the early stages of another breakthrough capable of pushing psychological forward. Namely, researchers have begun to "crowd source" experiments to obtain large amounts of data from many people in short order. Building off ingenious ideas, such as Luis von Ahn's "ESP Game" (a "game" that was the basis for how Google matches words to images; von Ahn & Dabbish, 2004), researchers have turned to outlets such as Amazon's Mechanical Turk (e.g., Buhrmester, Kwang, & Gosling, 2011) to rapidly distribute an experiment to many different participants.

Another outlet for rapid distribution of experiments is through the use of mobile applications—"apps" created for mobile devices, such as Apple and Android products. Psychologists have a history of using games and game-like interfaces to make experiments more palatable to participants (e.g., Anguera et al., 2013; Boot et al., 2010; Mané & Donchin, 1989; Miranda & Palmer, 2014), and mobile devices offer an exciting new means to crowd source an experiment in a game-like form.

Some mobile apps have been specifically designed to assess and/or train cognitive abilities, and they can address open questions with data voluntarily contributed by users. Other mobile apps just happen to tap into cognitive abilities in a manner that can be analyzed by researchers, even though that might not have been the apps' intended purpose. For example, some games challenge players to look for differences between images presented side-by-side—a game version of change detection tasks (e.g., Simons & Rensink, 2005). Similarly, other games tap into abilities related to the cognitive processes of working memory (memory match games), go/no-go ("whack-a-mole" games), and visual search (search-and-find games).

Using Mobile Technology for Research

There are clear advantages to crowd sourcing data collection through mobile technology. The most obvious benefit is the potential for gathering "big data"—massive datasets that provide the ability to examine nuanced questions with sufficient statistical power. Likewise, this can provide a means to collect relatively cheap data in an automated and continuous manner. Lastly, this process can mimic real-world aspects that are difficult to address in a laboratory environment (e.g., realistic distributions of variables).

There are also clear disadvantages to consider. First, researchers either need to have the necessary skills to create a fun game or need to partner with a developer. Gathering data through a mobile

This article was published Online First December 8, 2014.

Stephen R. Mitroff, Adam T. Biggs, Stephen H. Adamo, Emma Wu Dowd, Jonathan Winkle, and Kait Clark, Department of Psychology and Neuroscience, Center for Cognitive Neuroscience, Duke University.

We thank Ben Sharpe, Thomas Liljetoft, and Kedlin Company for access to the *Airport Scanner* data and for approving the use of *Airport Scanner* images.

Correspondence concerning this article should be addressed to Stephen R. Mitroff, Center for Cognitive Neuroscience, B203 LSRC, Box 90999, Duke University, Durham, NC 27708. E-mail: mitroff@duke.edu

game is only worthwhile if people will play the game, and people are more likely to play if the game is fun (e.g., Miranda & Palmer, 2014). Second, large amounts of data may not necessarily result in high-quality data; it is critical to carefully select what research questions are to be addressed and how they are addressed through the game interface. Finally, by collecting data through crowd sourcing, there is an inherent lack of control over who is playing and under what conditions (e.g., there is no way to know what percentage of the data is collected while participants are on the toilet).

Assuming the advantages outweigh the disadvantages and that the disadvantages can be addressed, the largest benefit of data collection through mobile technology is the potential for analyzing big data. With millions (or billions) of trials, it is possible to examine experimental variables that are too difficult to assess in a laboratory environment. While using mobile technology for research purposes may seem like a simple methodological advance, it has the potential to greatly inform psychology theory. Here we discuss our recent efforts focusing on the specific cognitive task of *visual search*.

Examples of Using Mobile Technology for Research

Visual search is the act of looking for target items among distractor items. Decades of research have sought to understand this ubiquitous cognitive process and to determine how humans, nonhuman animals, and computers successfully identify targets (see Eckstein, 2011; Horowitz, 2014; Nakayama & Martini, 2011, for recent reviews). Visual search has a history of using big data analyses—in 1998, Jeremy Wolfe collated data from 2,500 experimental sessions to ask "What can 1 million trials tell us about visual search" (Wolfe, 1998). This endeavor confirmed some open hypotheses and challenged others, while also demonstrating the value of big data for visual search analyses.

The downside of Wolfe's approach was that it took 10 years to collect—as is to be expected in typical laboratory experiments. Mobile apps offer the potential to collect data far more expeditiously. In the current report, we discuss results from our recent partnership with Kedlin Co., the creators of *Airport Scanner* (https://www.airportscannergame.com). In *Airport Scanner*, players are tasked with searching for illegal items in simulated x-ray bag images in an airport security environment. Players view one bag at a time and use finger-taps to identify illegal items on a touchscreen (see Figure 1 for gameplay examples). Players are provided with a logbook of illegal and legal items, and the logbook expands (going from a handful of possible targets to hundreds) as players progress through the game.

As of November 2014, there were over two billion trials from over seven million mobile devices available for research purposes. Data are collected in accordance with the terms and conditions of the standard Apple User Agreement and those provided by Kedlin Co. Each player consents to the terms and conditions when installing the application, and the Duke University Institutional Review Board provided approval for secondary data analyses (see Biggs, Adamo, & Mitroff, 2014; Mitroff & Biggs, 2014, for more details). Here we provide a brief overview of two examples of how we have used the *Airport Scanner* data for research purposes.

Use of Airport Scanner Data

Example 1: Ultra-rare Targets

In a recently published article (Mitroff & Biggs, 2014), we explored how visual search performance is affected when specific targets rarely appeared. While maintaining an overall target prevalence rate of 50% (half of the *bags* in *Airport Scanner* had at least one target present), the frequency with which any given target

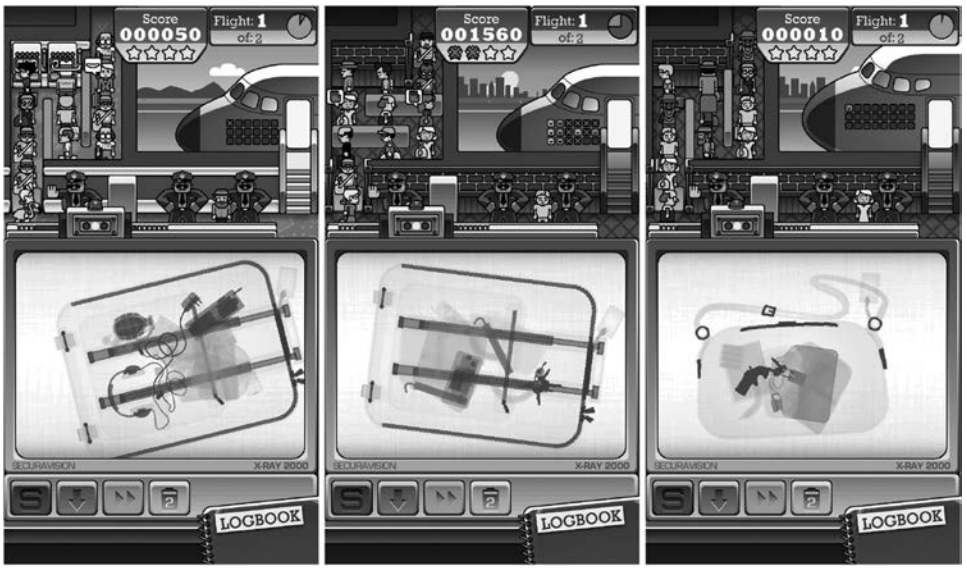

Figure 1. Sample images from *Airport Scanner*: the left image contains one target (*hand grenade*), the middle contains two identical targets (two exemplars of the *dynamite stick* target type), and the right image contains two different target types (*derringer*, *gasoline can*). *Airport Scanner* images appear with permission from Kedlin Co. Copyright 2014 by the Kedlin Company. See the online article for the color version of this figure.

could appear varied greatly (e.g., a *hammer* appeared as a target in 3% of the trials while a *switchblade* appeared as a target in only 0.08% of the trials). Critically, nearly 30 of the targets were "ultra-rare"—they appeared in less than 0.15% of all trials. To examine the effects of such extreme target rarity on visual search performance in a laboratory would be difficult for even one target item. For example, to assess accuracy for targets that only appeared in 0.1% of trials, 1,000 trials would be needed for a single occurrence. To obtain sufficient statistical power (e.g., at least 20 occurrences), too many total trials would be needed to realistically test such a question in a laboratory. However, with the large *Airport Scanner* dataset, we were able to look at hundreds of cases for each of the nearly 30 "ultra-rare" targets. Comparing the relationship between search accuracy and target frequency across 78 unique target types of various frequency rates revealed an extremely strong logarithmic relationship (adjusted $R^2 = .92$) such that the "ultra-rare" items were much more likely to be missed than the more frequently occurring targets (Mitroff & Biggs, 2014). This example highlights the more obvious benefits and drawbacks of using big data to address research questions. The primary benefit is clear—a question that could have taken decades to answer in a laboratory setting can be answered using big data in a fraction of the time. The *Airport Scanner* app also bypasses most of the potential downsides mentioned above given that it is a popular game with an interface that is conducive to research. However, there is an inherent lack of control over the nature of data collected via mobile technology, and there is no obvious means to counter this lack of control. Analogous to a "speed/accuracy trade-off" (a well-studied juxtaposition between performing quickly vs. performing accurately; e.g., Pachella, 1974), big data might engender a "volume/control trade-off"—a juxtaposition between the amount of available data and the methodological control over the data.

Example 2: Multiple-Target Search Theories

Many real-world visual searches can have more than one target present within the same search array (e.g., more than one abnormality in a radiological x-ray; more than one prohibited item in a bag going through airport security). Unfortunately, multiple-target searches are highly susceptible to errors such that additional targets are less likely to be detected if one target has already been found (see Berbaum, 2012, for a review). This effect was originally termed the "satisfaction of search" phenomenon, but we have recently renamed it the "subsequent search misses" (SSM) phenomenon (Adamo, Cain, & Mitroff, 2013). SSM errors are a stubborn source of errors, and several efforts (e.g., Berbaum, 2012; Cain, Adamo, & Mitroff, 2013) have attempted to identify their underlying cause(s).

Three primary theories of SSM have been proposed. First, the original explanation—and the source of the "satisfaction of search" name—suggests that searchers become "satisfied" with the meaning of the search on locating a first target and terminate their search prematurely (Smith, 1967; Tuddenham, 1962). Second, a resource depletion account (e.g., Berbaum et al., 1991; Cain & Mitroff, 2013) suggests that cognitive resources (e.g., attention, working memory) are consumed by a found target and leave fewer resources available to detect additional targets during subsequent search. Finally, a perceptual set account suggests that searchers become biased to look for additional targets similar to the first found target (Berbaum et al., 1990; Berbaum et al., 1991; Berbaum et al., 2010); for example, if you just found a tumor, you might enter "tumor mode" and be less likely to subsequently detect a fracture that appeared in the same x-ray image.

There have been empirical tests of the satisfaction and resource depletion theories (see Berbaum, 2012), but no substantial tests have been offered for the perceptual set account. Previous investigations have employed a small number of possible target types; for example, Fleck, Samei, and Mitroff (2010) asked observers to search for targets that were T-shaped items among L-shaped distractor items. There were two different forms of the target Ts—those that were relatively light and those that were relatively dark. If a perceptual set operates via a priming-like influence, such a design might be suboptimal, because repeated exposure to each target could result in elevated priming across all trials for all targets (i.e., you only need to see so many lightly colored T-shaped targets before you become generally biased for lightly colored T-shaped targets across the entire experiment).

A large and unpredictable set of targets could generate more short-lived priming during a visual search task, which is more in line with real-world scenarios in which a perceptual set could meaningfully impact performance. However, such an experimental design—many and varied targets spread over an immense number of trials—is not practical to administer in a laboratory environment, and it is not easily assessed in real-world scenarios such as an airport security checkpoint. Here we used the *Airport Scanner* gameplay data to address this idea.

More details about the nature of the data and the gameplay are available elsewhere (e.g., Mitroff & Biggs, 2014); however, we highlight in Table 1 how we filtered the gameplay variables to address our specific research question at hand. Each trial (*bag*) could contain 0–3 illegal target items with approximately 43% of all trials containing one target, 6% with two targets, and less than 1% with three targets. We analyzed SSM errors by comparing search accuracy for a specific target item on single-target trials to search accuracy for the same target on dual-target trials when another target had been detected first (e.g., Biggs & Mitroff, 2014). Players identified a target by tapping directly onto the target location, and we excluded all cases in which one tap captured two targets. Analyses were limited to target types with at least 20 instances within our dataset (i.e., the *shotgun* target type was filtered from our SSM analyses for only contributing seven instances).

We first determined whether SSM errors occurred in the *Airport Scanner* gameplay as we have observed in the laboratory (e.g., Cain et al., 2013; Fleck et al., 2010). This analysis was performed across 78 target items without considering the identity of the first found target; for example, when calculating the SSM error rate for the *pistol* as a second target, all data were included whether or not the first found target was another *pistol*, a *grenade*, a *knife*, and so forth. A significant overall SSM error rate was found ($M = 14.00\%$, $SE = 1.11\%$), $t(77) = 12.62$, $p < .001$.

Next, we assessed SSM errors when the two targets in the same bag were identical (e.g., two exemplars of the *dynamite stick* target type; e.g., the second panel of Figure 1) as opposed to when the two targets in the same bag were not identical (e.g., one *pistol* and one *hand grenade*; e.g., the third panel of Figure 1). Thirty-three target types met the 20-occurrence minimum for inclusion into

Table 1
Game Elements and Nature of Trial Filtering for the Multiple-Target Visual Search Example

Game element/variable			
Game variable	Description	Cases	Filtering for SSM errors example
Airport	6 levels; increase in difficulty	*Trainee*, *Honolulu*, *Las Vegas*, *Chicago*, *Aspen*, *London*	Exclude *Trainee*
Rank	5 levels; player's experience level	*Trainee*, *Operator*, *Pro*, *Expert*, *Elite*	Only *Elite* players (here = 62,606 devices)
Day	Sessions within Airport level	5 *Days* per *Airport*; additional *Challenge* levels for some *Airports*	Exclude *Challenge*
Mission type	Game play mode	*Career*, *Challenge*	Only *Career* mode
Replay	Repeat a Day after completing it	Replays allowed or disallowed	Replays allowed
Day status	How the Day session ended	Completed, timed out, security breach	No exclusions
Bag type	Shape and size of search array	> 15 unique *Bag* types	*Briefcase*, *carry-on*, *duffle*, and *purse* included
Passenger type	Difficulty of Bag	*Easy*: ≈ 0–8 legal items present *Medium*: ≈ 9–13 legal items present *Hard*: ≈ 14–20 legal items present *Impossible*: Requires upgrades	Exclude *Impossible*
In-game upgrades	Add-ons to help with gameplay	> 10 unique upgrades	Exclude all that affect search performance
Special passengers/items	Nontypical gameplay events	*Air Marshals*, *Flight Crew*, *First Class Passengers*, *Delay Passengers*, *Rare Targets (special items)*	*Air Marshal* and *rare-target Bags* excluded
Illegal item count	No. of target items present	0 targets: ≈ 50% of all trials 1 target: ≈ 43% of all trials 2 targets: ≈ 6% of all trials 3 targets: ≈ 1% of all trials	Excluded 0-illegal and 3-illlegal item *Bags*
Legal item count	No. of distractor items present	0–20	No exclusions
Specific illegal items	Various target items (see Figure 1)	> 200	1-target accuracy: $N = 78$ 2-target accuracy: $N = 33$
Specific legal items	Various distractor items (see Figure 1)	> 200	No exclusions

How data filtering affected trial counts	
Total trials available as of 11/18/14	2,236,844,667
Total trials for example analysis date range of 04/15/13 to 08/26/13	1,098,098,764
Total trials for 1-target trial accuracy analyses after all filters applied	1,795,907
Total trials for 2-target trial accuracy analyses after all filters applied	126,579

Note. SSM = subsequent search misses.

analyses. We observed significant SSM errors when the first and second targets were identical ($M = 6.53\%$, $SE = 1.62\%$), $t(32) = 4.04$, $p < .001$, and when the first and second targets were not identical ($M = 19.21\%$, $SE = 1.36\%$), $t(32) = 14.13$, $p < .001$. Importantly, there was a significant difference in SSM error rates for identical targets versus nonidentical targets ($M_{\text{difference}} = 12.69\%$, $SE = 1.69\%$), $t(32) = 7.52$, $p < .001$, such that SSM errors were substantially reduced when the targets were identical than when the two targets were not identical.

In this particular example, the power of big data allowed us to answer a nuanced question that required a substantial number of trials to appropriately assess. Specifically, this analysis focused on trials that contained two targets and that met several exclusionary criteria (see Table 1), which resulted in analyzing only about 125,000 trials out of a dataset of more than 1,000,000,000 trials (0.01%). It was necessary to examine accuracy for specific target types within a framework containing a wide variety of target types to prevent participants from becoming overexposed to any one target. This was best accomplished through the use of mobile technology, which allowed for the accumulation of the necessary data while providing players an enjoyable experience. Importantly, we expanded current understanding about the mechanisms underlying SSM errors by revealing that the errors are, in fact, partially due to a perceptual set mechanism.

Conclusion

Using a game interface to assess cognitive abilities is not new to psychological research (e.g., Boot et al., 2010; Mané, & Donchin, 1989), but mobile technology offers a phenomenal opportunity to examine cognitive processes on a large scale. Here we discussed two specific examples of how we analyzed data from the *Airport Scanner* game to address psychological questions, and much more is possible.

However, using mobile apps for research purposes is easier said than done. Researchers can build games for data collection purposes and have complete control over the design and implementation. However, there is no guarantee that any game will be successful enough to garner data—simply producing a game does not ensure anyone will play it. Alternatively, researchers can partner with developers to create apps or can partner with devel-

opers of already existing apps, which can be a great opportunity for both groups; developers can benefit from researchers' insight and added press, and researchers can benefit from developers' skill and access to established games. Our partnership with Kedlin Co. exemplifies this beneficial relationship and has been successful enough to lead to federal funding opportunities for further research implementations of *Airport Scanner*.

With the proliferation of mobile technology, it is time to aim high. In 1998, 1 million trials on a specific cognitive task was a mind-blowing amount of data (Wolfe, 1998). Today, we collect over a million trials a day through *Airport Scanner*. Inflation has hit visual search, and as researchers, we should (responsibly and carefully) look for ways to take advantage of this opportunity.

References

Adamo, S. H., Cain, M. S., & Mitroff, S. R. (2013). Self-induced attentional blink: A cause of errors in multiple-target search. *Psychological Science, 24,* 2569–2574. http://dx.doi.org/10.1177/0956797613497970

Anguera, J. A., Boccanfuso, J., Rintoul, J. L., Al-Hashimi, O., Faraji, F., Janowich, J., . . . Gazzaley, A. (2013). Video game training enhances cognitive control in older adults. *Nature, 501,* 97–101. http://dx.doi.org/10.1038/nature12486

Berbaum, K. S. (2012). Satisfaction of search experiments in advanced imaging. *Proceedings of the Society for Photo-Instrumentation Engineers, 8291,* 82910V. http://dx.doi.org/10.1117/12.916461

Berbaum, K. S., Franken, E. A., Jr., Dorfman, D. D., Rooholamini, S. A., Coffman, C. E., Cornell, S. H., . . . Smith, T. P. (1991). Time course of satisfaction of search. *Investigative Radiology, 26,* 640–648. http://dx.doi.org/10.1097/00004424-199107000-00003

Berbaum, K. S., Franken, E. A., Jr., Dorfman, D. D., Rooholamini, S. A., Kathol, M. H., Barloon, T. J., . . . Montgomery, W. J. (1990). Satisfaction of search in diagnostic radiology. *Investigative Radiology, 25,* 133–140. http://dx.doi.org/10.1097/00004424-199002000-00006

Berbaum, K. S., Franklin, E. A., Jr., Caldwell, R. T., & Schartz, K. M. (2010). Satisfaction of search in traditional radiographic imaging. In E. Samei & E. Krupinski (Eds.), *The handbook of medical image perception and techniques* (pp. 107–138). New York, NY: Cambridge University Press.

Biggs, A. T., Adamo, S. H., & Mitroff, S. R. (2014). Rare, but obviously there: Effects of target frequency and salience on visual search accuracy. *Acta Psychologica, 152,* 158–165. http://dx.doi.org/10.1016/j.actpsy.2014.08.005

Biggs, A. T., & Mitroff, S. R. (2014). Different predictors of multiple-target search accuracy between nonprofessional and professional visual searchers. *The Quarterly Journal of Experimental Psychology, 67,* 1335–1348. http://dx.doi.org/10.1080/17470218.2013.859715

Boot, W. R., Basak, C., Erickson, K. I., Neider, M., Simons, D. J., Fabiani, M., . . . Kramer, A. F. (2010). Transfer of skill engendered by complex task training under conditions of variable priority. *Acta Psychologica, 135,* 349–357. http://dx.doi.org/10.1016/j.actpsy.2010.09.005

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6,* 3–5. http://dx.doi.org/10.1177/1745691610393980

Cain, M. S., Adamo, S. H., & Mitroff, S. R. (2013). A taxonomy of errors in multiple-target visual search. *Visual Cognition, 21,* 899–921. http://dx.doi.org/10.1080/13506285.2013.843627

Cain, M. S., & Mitroff, S. R. (2013). Memory for found targets interferes with subsequent performance in multiple-target visual search. *Journal of Experimental Psychology: Human Perception and Performance, 39,* 1398–1408. http://dx.doi.org/10.1037/a0030726

Eckstein, M. P. (2011). Visual search: A retrospective. *Journal of Vision, 11*(5), 14. http://dx.doi.org/10.1167/11.5.14

Fleck, M. S., Samei, E., & Mitroff, S. R. (2010). Generalized "satisfaction of search": Adverse influences on dual-target search accuracy. *Journal of Experimental Psychology: Applied, 16,* 60–71. http://dx.doi.org/10.1037/a0018629

Horowitz, T. S. (2014). Exit strategies: Visual search and the quitting time problem. In R. Metzler, G. Oshanin, & S. Redner (Eds.), *First-passage phenomena and their applications* (pp. 390–415). Hackensack, NJ: World Scientific Press.

Mané, A., & Donchin, E. (1989). The Space Fortress game. *Acta Psychologica, 71,* 17–22. http://dx.doi.org/10.1016/0001-6918(89)90003-6

Miranda, A. T., & Palmer, E. M. (2014). Intrinsic motivation and attentional capture from gamelike features in a visual search task. *Behavior Research Methods, 46,* 159–172. http://dx.doi.org/10.3758/s13428-013-0357-7

Mitroff, S. R., & Biggs, A. T. (2014). The ultra-rare-item effect: Visual search for exceedingly rare items is highly susceptible to error. *Psychological Science, 25,* 284–289. http://dx.doi.org/10.1177/0956797613504221

Nakayama, K., & Martini, P. (2011). Situating visual search. *Vision Research, 51,* 1526–1537. http://dx.doi.org/10.1016/j.visres.2010.09.003

Pachella, R. (1974). The interpretation of reaction time in information processing research. In B. H. Kantowitz (Ed.), *Human information processing: Tutorials in performance and cognition* (pp. 41–82). Hillsdale, NJ: Erlbaum.

Simons, D. J., & Rensink, R. A. (2005). Change blindness: Past, present, and future. *Trends in Cognitive Sciences, 9,* 16–20. http://dx.doi.org/10.1016/j.tics.2004.11.006

Smith, M. J. (1967). *Error and variation in diagnostic radiology*. Springfield, IL: C. C. Thomas.

Tuddenham, W. J. (1962). Visual search, image organization, and reader error in roentgen diagnosis. Studies of the psychophysiology of roentgen image perception. *Radiology, 78,* 694–704. http://dx.doi.org/10.1148/78.5.694

von Ahn, L., & Dabbish, L. (2004). Labeling images with a computer game. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* 319–326. http://dx.doi.org/10.1145/985692.985733

Wolfe, J. M. (1998). What can 1 million trials tell us about visual search? *Psychological Science, 9,* 33–39. http://dx.doi.org/10.1111/1467-9280.00006

Received June 10, 2014
Revision received September 2, 2014
Accepted September 6, 2014 ■