ASSOCIATION FOR
PSYCHOLOGICAL SCIENCE

Psychological Science
2014, Vol. 25(1) 284–289
© The Author(s) 2013
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0956797613504221
pss.sagepub.com
SAGE

The Ultra-Rare-Item Effect: Visual Search for Exceedingly Rare Items Is Highly Susceptible to Error

Stephen R. Mitroff and Adam T. Biggs
Department of Psychology and Neuroscience, Center for Cognitive Neuroscience,
Duke University

Abstract

Accuracy is paramount in radiology and security screening, yet many factors undermine success. Target prevalence is a particularly worrisome factor, as targets are rarely present (e.g., the cancer rate in mammography is ~0.5%), and low target prevalence has been linked to increased search errors. More troubling is the fact that specific target types can have extraordinarily low frequency rates (e.g., architectural distortions in mammography—a specific marker of potential cancer—appear in fewer than 0.05% of cases). By assessing search performance across millions of trials from the *Airport Scanner* smartphone application, we demonstrated that the detection of ultra-rare items was disturbingly poor. A logarithmic relationship between target detection and target frequency (adjusted $R^2 = .92$) revealed that ultra-rare items had catastrophically low detection rates relative to targets with higher frequencies. Extraordinarily low search performance for these extraordinarily rare targets—what we term the *ultra-rare-item effect*—is troubling given that radiological and security-screening searches are primarily ultra-rare-item searches.

Keywords
visual search, visual attention, visual memory

Received 5/20/13; Revision accepted 8/13/13

In radiological exams and security screenings, it is vital to find all targets, as missed tumors or contraband could have life-threatening implications. However, examining the nature of such real-world searches can be difficult because they can vary greatly from standard laboratory search tasks (Clark, Cain, Adamo, & Mitroff, 2012). For example, in most laboratory search experiments, a target is present in half of the displays, but targets are seldom present in radiological exams and security screenings. It is estimated that the cancer rate in mammography is 4.69 cancers per 1,000 examinations (~0.5% of cases examined; Breast Cancer Surveillance Consortium, 2009). Moreover, radiologists often search for specific cancerous features that can be exceedingly rare; for example, in an examination of 300 consecutive screening-detected breast cancers (Sickles, 1986), only 9% showed architectural distortions (an abnormal arrangement of tissue that suggests the presence of cancer). Extrapolating these probabilities suggests that architectural distortions will be present in only 0.042% of mammography screenings.

Previous research has found that visual search is disturbingly inaccurate when targets are rarely present—a phenomenon termed the *low-prevalence effect* (Evans, Evered, Tambouret, Wilbur, & Wolfe, 2011; Wolfe, Horowitz, & Kenner, 2005). Searchers may adjust their criterion when targets are seldom found (Gur et al., 2007; Wolfe et al., 2005; but see Fleck & Mitroff, 2007; Gur et al., 2003), which reduces the likelihood of finding targets that are rarely present. The low-prevalence effect is driven by the percentage of cases that contain a target, which raises concerns for many real-world searches. However, another potential cause of error has gone unstudied: Regardless of overall target prevalence, is a target type that is exceedingly rare (e.g., architectural distortions in cancer screenings) at particular risk to be

Corresponding Author:
Stephen R. Mitroff, LSRC Building, Room B249, Box 90999, Duke University, Durham, NC 27708
E-mail: mitroff@duke.edu

missed? To appropriately distinguish between these two potential influences on visual search accuracy, we use the term *prevalence* to denote overall target probability (i.e., the percentage of trials containing a target, regardless of which target), and we use the word *frequency* to denote a particular target item's rate of appearance (i.e., the percentage of trials containing that specific target).

Previous studies (Wolfe et al., 2005; Wolfe et al., 2007) assessed search accuracy with overall target prevalence at or above 50%, but with one particular class of targets appearing on only 1% of trials. The rare target type had a higher miss rate than the others, which suggests a focused effect of frequency for specific targets. Yet 1% frequency is quite high in practical scenarios; although contraband may be present in 1% or more of the bags at security checkpoints, the frequency of any one class of items (e.g., guns) can be an order of magnitude lower. Assessing search performance for targets with frequency rates below 1% is critical given that most real-world searches include targets with ultra-low frequencies, but this presents an intractable problem for laboratory-based studies because ultra-rare frequency levels require an enormous number of trials. To address this important issue, we examined search accuracy for multiple ultra-rare items using the *Airport Scanner* smartphone application (Kedlin Co., www.airportscannergame.com).

In *Airport Scanner*, players serve as airport security-checkpoint officers and search for illegal items in bags (Fig. 1). We assessed anonymous game-play data from thousands of players and from nearly 6 million trials in the primary data set. Here we present data from trials involving 78 unique targets with frequency rates as low as 0.078%. Overall target prevalence was 50% (i.e., half the searches had at least one target present), but the probability of a given target being present could be ultra-rare, as 30 items had rates below 0.15%.

Method

Analyses were conducted on anonymous *Airport Scanner* game-play data that were recorded in accordance with the standard Apple User Agreement; data use was approved by Duke University's institutional review board. All data were drawn from players who voluntarily installed the application and played the game between December 1, 2012, and March 18, 2013.

Airport Scanner comprises multiple levels (airports); each level involves multiple sessions (days), and each session contains multiple trials (bags). The bags vary in size, shape, orientation, number of legal items (i.e., number of distractors), and number of illegal items. The 78 illegal items that served as targets in the current analyses are the illegal items that appeared in the briefcase and carry-on bags at the Honolulu and Las Vegas airport levels. During game play, players advance through five status ranks, with *elite* being the highest. We focused our analyses on elite players because these players had already completed the Honolulu and Las Vegas levels at a lower status (i.e., the current data were from trials on which the players replayed the levels after achieving elite status), and because they had prior exposure to the 78 illegal items. In this way, we eliminated concerns over the time course of new items being introduced into game play and how that might influence target frequency. (See the Supplemental Material available online for more information about the players and the game play.)

The primary dependent measure was hit rate—the percentage of correct detections of a given target—and the primary independent measure was item frequency—the percentage of bags (trials) containing each specific target. Hit rates for each of the 78 unique targets were calculated from 11,053 players and 370,468 bags. These data come from a highly constrained data set, so that very little varied across trials except the frequency of the targets. For example, all trials contributing to these calculations had exactly one illegal item present in the bag, and the bags were of only two possible types (carry-on or briefcase; see the Supplemental Material for full parameters).

Item frequency was calculated from a larger data pool that represents the players' exposure to the target items. Frequency rates were calculated from 5,847,292 bags that contained a total of 3,333,181 targets (each bag had 0 to

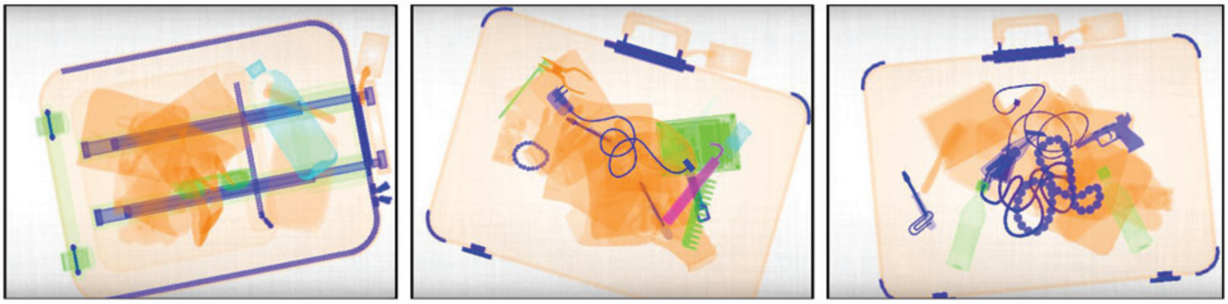

Fig. 1. Sample trials from *Airport Scanner* (presented with permission from Kedlin Co.). Each sample bag contains one illegal target item (from left to right, a big bottle, dynamite stick, and gun part) and numerous legal (nontarget) items.

20 nontargets and 0, 1, 2, or 3 targets). These trials maintained the global constraints from the hit rate data (data from the Honolulu and Las Vegas airports only, from elite players only, from the same date range, etc.) but included the full range of targets experienced (0-, 1-, 2-, and 3-target bags, all bag types, in-game upgrades allowed,[1] etc.). Targets ranged in frequency from 0.078% (4,583/5,847,292) to 3.722% (217,609/5,847,292). (See the Supplemental Material for more details regarding target frequencies.)

Results

Overall hit and false alarm rates were 89.84% and 11.01%, respectively. We conducted a nonlinear regression analysis and observed a strong logarithmic relationship (Wolfe, 2012) between hit rate and frequency (natural logarithm function), adjusted (Adj.) $R^2 = .70$, $F(1, 76) = 176.82$, $p < .001$; ultra-rare items were highly likely to be missed, whereas more frequently present items were easily detected (Fig. 2). The data were also significantly explained by linear and quadratic functions (Adj. R^2s = .50 and .66, respectively), but the logarithmic function provided the best fit. The 30 targets with a frequency rate below 0.15% were detected only 27% of the time, whereas the 23 targets with rates above 1% were detected on 92% of the trials. Aggregate data, in which targets with highly similar frequency rates were combined to minimize hit rate variability based on single-target rates, confirmed the relationship between hit rate and frequency (linear: Adj. $R^2 = .53$; quadratic: Adj. $R^2 = .86$, logarithmic: Adj. $R^2 = .92$; all *p*s < .001; Fig. 3).

Frequency versus salience

In *Airport Scanner*, more frequent targets are typically more visually salient. Some targets are visually distinct and relatively easy to detect (e.g., they are relatively large), and several of these targets also appear more frequently than other targets. Thus, it is important to rule out salience-based explanations for the relation between hit rate and frequency. Because the Honolulu and Las Vegas airports differed little except in how frequently certain targets appeared, we were able to address this alternative explanation by comparing the two airports' hit rate data. We found that a target was more likely to be detected when it appeared more frequently at a given airport: The 10 targets appearing more frequently in Honolulu than in Las Vegas, $t(9) = 14.89$, $p < .001$ (paired-sample two-tailed test), had a higher hit rate in Honolulu, $t(9) = 9.80$, $p < .001$, and the 10 targets appearing more frequently in Las Vegas than in Honolulu, $t(9) = 8.11$, $p < .001$, had a higher hit rate in Las Vegas, $t(9) = 3.36$, $p = .01$. Moreover, the hit rate for the 10 targets that appeared most equivalently at the two airports, $t(9) = 0.70$, $p = .50$, did not differ between the airports, $t(9) = 0.55$, $p = .60$.

Frequency versus visual clutter

For the bags used to calculate hit rates, the number of legal items (nontarget distractors) ranged from 1 to 19. Given this variability, there may be concern that the observed relation between hit rate and frequency was skewed by clutter: If the less frequent items were disproportionately more difficult to detect in cluttered bags, then overall bag difficulty could possibly account for the results. To address such potential concerns, we assessed the relation between hit rate and frequency only for trials with at least 14 nontargets present (i.e., the most cluttered bags; $n = 40{,}159$). The logarithmic relationship between hit rate and frequency held even for these highly cluttered search arrays (Adj. $R^2 = .67$, $p < .001$).

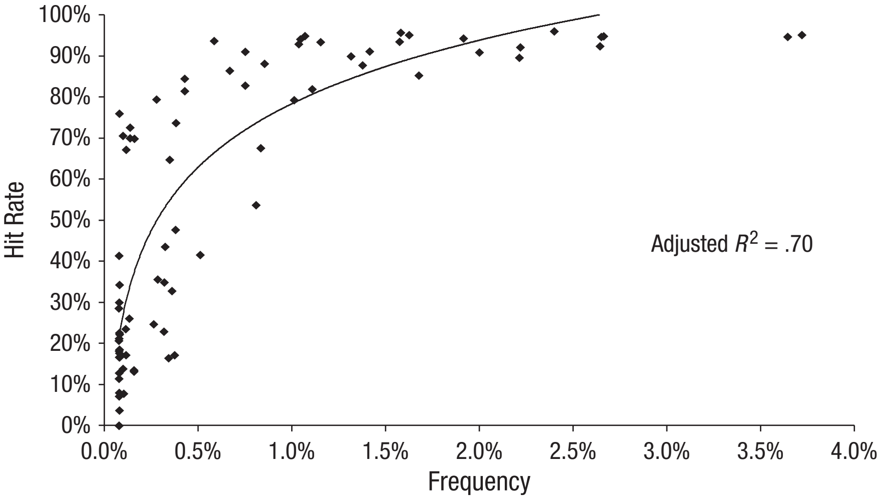

Fig. 2. Hit rate as a function of frequency for the 78 unique target types. The adjusted R^2 value represents the fit of the logarithmic function (solid line).

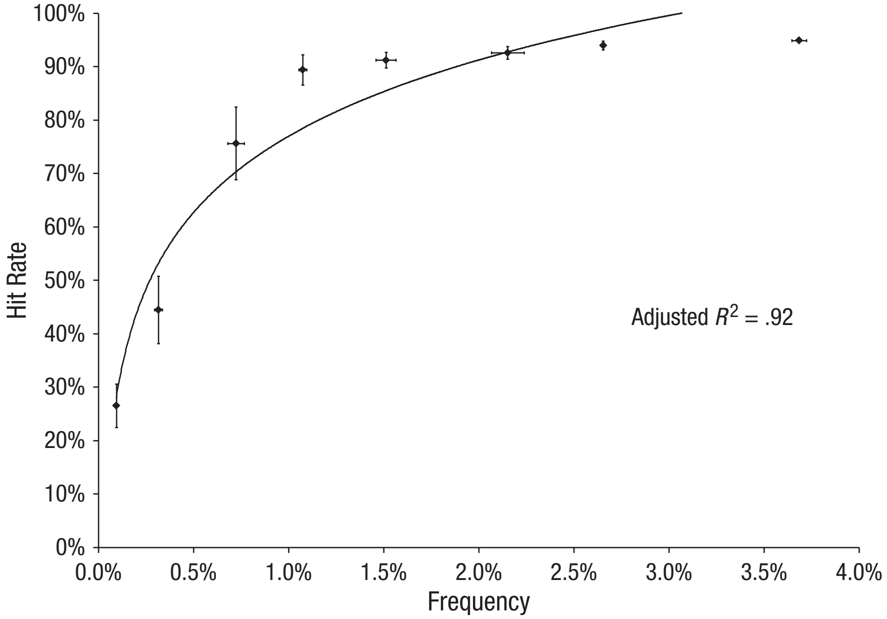

Fig. 3. Hit rate as a function of frequency for aggregated data. Each point represents results for targets with similar frequency rates. The adjusted R^2 value represents the fit of the logarithmic function (solid line). Horizontal error bars represent standard errors of the mean for frequency, and vertical error bars represent standard errors of the mean for hit rate.

Five independent replications

We implemented specific filters for calculating the hit rate data to reduce variability, eliminate extraneous factors above and beyond frequency, and allow for the salience comparison between two airports. Might the observed relationship between hit rate and frequency be due to something particular about this selected set? The nature of the *Airport Scanner* data set allowed us to address such concerns by replicating the effect with nonoverlapping data. Table 1 summarizes five independent replications, which minimize concerns that specific parameters possibly influenced the results because the same relationship between hit rate and frequency was found. The six nonoverlapping data sets (the primary analysis and the five replications) showing this relationship collectively included hit rate data from more than 2 million trials and frequency data from more than 17 million trials.

Discussion

Ultra-rare items are highly vulnerable to being missed in visual search; in a task in which relatively frequent items were detected more than 90% of the time, ultra-rare items (those with frequency rates below 0.15%) were largely undetected. We refer to this newly demonstrated constraint on visual search accuracy as the ultra-rare-item effect. This effect is driven by the frequency of specific target items, not the overall prevalence of targets during the search task (i.e., the low-prevalence effect; Wolfe et al., 2005), or by target salience or visual clutter.

In this study, the ultra-rare-item effect manifested itself as a nonlinear influence of frequency on search accuracy: Accuracy was considerably worse for targets with frequency rates below 1% than for targets with frequency rates above 1% (Figs. 2 and 3). Previous research on the low-prevalence effect (e.g., Wolfe et al., 2007) is consistent with the logarithmic relationship found here, and the current results solidify this trend because they are based on data from a large number of target items. In practice, the observed nonlinear function describes a sharp cliff in accuracy, such that there is grave danger of targets with frequencies below 1% being missed.

Target items were present in 50% of the search arrays included in the current analyses, which eliminates several related, but orthogonal, potential explanations of the findings. Specifically, because participants were actively finding targets on a regular basis, the observed accuracy effects could not have been driven by a global criterion shift (e.g., Menneer, Donnelly, Godwin, & Cave, 2010; Wolfe & Van Wert, 2010), a vigilance decrement (Mackworth, 1970), or the low-prevalence effect (Wolfe et al., 2005). Likewise, participants did not develop a prepotent motor response bias that could have resulted in simple motor errors driving accuracy deficits (e.g., Fleck & Mitroff, 2007).

Table 1. Summary of Five Independent Replications of the Relationship Between Hit Rate and Target Frequency

Replication	Player rank	Airports	Bag types	Hit rate trials[a]	Frequency trials[a]	Adjusted R^2
1	Elite	Chicago, Aspen, London	Carry-on, briefcase	229,596	1,643,207	.47
2	Elite	Honolulu, Las Vegas	Duffel bag, purse	168,285	5,847,292	.63
3	Expert	Honolulu, Las Vegas	Carry-on, briefcase	444,968	3,186,467	.46
4	Pro	Honolulu	Carry-on, briefcase	567,692	3,210,830	.77
5	Pro	Las Vegas	Carry-on, briefcase	222,287	3,309,035	.57

Note: The table presents adjusted R^2 values for a logarithmic relationship between hit rate and frequency, all *p*s < .001.
[a]These columns report the numbers of trials included in calculating hit rate and target frequency, respectively.

The ultra-rare-item effect may be driven by specific search expectations, an attentional set, or both (Berbaum, 2012; Leber & Egeth, 2006): Searchers might be "tuned" for detecting items that are more likely than others to be present. These data suggest a very fine-tuned system, as all 78 targets had a frequency rate below 4%. Target frequency may have both a short-term effect on accuracy by priming representations of the relatively frequent items and a long-term effect on accuracy as searchers adjust their expectations and search goals to minimize efforts toward finding ultra-rare items (Cain, Vul, Clark, & Mitroff, 2012).

The low-prevalence effect (Wolfe et al., 2005) can be partially attenuated by providing searchers with a short burst of high-prevalence search (Wolfe et al., 2007), which is thought to "reset" the searchers' criterion to a more effective state. Similarly, a short burst of ultra-rare items will not alter an item's overall frequency, but might have a lasting effect that could alter expectations, attentional priming, or both. The U.S. Transportation Security Administration can implement such a solution within the Threat Image Projection program, in which threatening items are projected onto passengers' bags at airport checkpoints (Hofer & Schwaninger, 2005). The relative rate of projected images could be strategically manipulated to provide security officers with short bursts of exposure to specific ultra-rare items that are deemed the most dangerous, in order to counteract the ultra-rare-item effect. Radiological search is moving to a mostly digital format and therefore can adopt a similar system of artificially altering frequency rates for particularly troublesome items.

Author Contributions

S. R. Mitroff and A. T. Biggs both contributed to analyzing the data, interpreting the data, and writing the manuscript.

Acknowledgments

We thank the Duke Visual Cognition lab for helpful comments during manuscript preparation, and Ben Sharpe, Thomas Liljetoft, and Kedlin Company for access to the *Airport Scanner* data.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Supplemental Material

Additional supporting information may be found at http://pss.sagepub.com/content/by/supplemental-data

Note

1. Players could obtain and activate in-game "upgrades," which helped game play (e.g., slowed down the conveyor belt to provide longer search times, or lessened the number of nontarget items in bags). When assessing search accuracy, we omitted trials in which a helpful upgrade was activated, to focus on unaided search performance. However, when assessing target frequency, we included trials with upgrades active, to obtain a full representation of how often each target item appeared during game play.

References

Berbaum, K. S. (2012). Satisfaction of search experiments in advanced imaging. *Proceedings of SPIE, 8291(1), Article 82910V*. Retrieved from http://proceedings.spiedigitallibrary.org/proceeding.aspx?articleid=1283355

Breast Cancer Surveillance Consortium. (2009). *Cancer rate (per 1,000 examinations) and cancer detection rate (per 1,000 examinations) for 1,960,150 screening mammography examinations from 2002 to 2006 by age*. Retrieved from http://breastscreening.cancer.gov/data/performance/screening/2009/rate_age.html

Cain, M. S., Vul, E., Clark, K., & Mitroff, S. R. (2012). A Bayesian optimal foraging model of human visual search. *Psychological Science*, *23*, 1047–1054.

Clark, K., Cain, M. S., Adamo, S. H., & Mitroff, S. R. (2012). Overcoming hurdles in translating visual search research between the lab and the field. In M. D. Dodd & J. H. Flowers (Eds.), *Nebraska Symposium on Motivation: Vol. 59. The influence of attention, learning, and motivation on visual search* (pp. 147–181). New York, NY: Springer.

Evans, K. K., Evered, A., Tambouret, R. H., Wilbur, D. C., & Wolfe, J. M. (2011). Prevalence of abnormalities influences cytologists' error rates in screening for cervical cancer.

Archives of Pathology & Laboratory Medicine, *135*, 1557–1560.

Fleck, M. S., & Mitroff, S. R. (2007). Rare targets are rarely missed in correctable search. *Psychological Science*, *18*, 943–947.

Gur, D., Bandos, A. I., Fuhrman, C. R., Klym, A. H., King, J. L., & Rockette, H. E. (2007). Prevalence effect in a laboratory environment: Changing the confidence ratings. *Academic Radiology*, *14*, 49–53.

Gur, D., Rockette, H. E., Armfield, D. R., Blachar, A., Bogan, J. K., Brancatelli, G., . . . Warfel, T. E. (2003). Prevalence effect in a laboratory environment. *Radiology*, *228*, 10–14.

Hofer, F., & Schwaninger, A. (2005). Using threat image projection data for assessing individual screener performance. *WIT Transactions on the Built Environment*, *82*, 417–426.

Leber, A. B., & Egeth, H. E. (2006). Attention on autopilot: Past experience and attentional set. *Visual Cognition*, *14*, 565–583.

Mackworth, J. (1970). *Vigilance and attention*. Harmondsworth, England: Penguin.

Menneer, T., Donnelly, N., Godwin, H. J., & Cave, K. R. (2010). High or low target prevalence increases the dual-target cost in visual search. *Journal of Experimental Psychology: Applied*, *16*, 122–144.

Sickles, E. A. (1986). Mammographic features of 300 consecutive nonpalpable breast cancers. *American Journal of Roentgenology*, *146*, 661–663.

Wolfe, J. M. (2012). Saved by a log: How do humans perform hybrid visual and memory search? *Psychological Science*, *23*, 698–703.

Wolfe, J. M., Horowitz, T. S., & Kenner, N. M. (2005). Rare items often missed in visual searches. *Nature*, *435*, 439–440.

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

Wolfe, J. M., & Van Wert, M. (2010). Varying target prevalence reveals two dissociable decision criteria in visual search. *Current Biology*, *20*, 121–124.