TITLE: PROJECT 2 EXAMPLE: Feedback X Prevalence Using Dermatology Stimuli

NCT#: **NCT05244122**

Date: 9.26.22

This is a composite document. The IRB approval for this study fall under a general approval that we use for many minimal risk behavioral studies. Under the revised Common Rule, it is unclear that we actually need IRB review.

This PDF concatenates

1) The protocol for "Prevalence Effects in Visual Search: Theoretical and Practical Implications"
2) A much more specific study protocol, edited text from Wolfe, J. M. (2022). How one block of trials influences the next: Persistent effects of disease prevalence and feedback on decisions about images of skin lesions in a large online study. *Cognitive Research: Principles and Implications (CRPI), 7*, 10. doi: https://doi.org/10.1186/s41235-022-00362-0

Prevalence effects in visual search: Theoretical and practical implications
Detailed Protocol 2022

Version: Prevalence_AUGUST2022
Version Date: AUGUST 12, 2022

I. BACKGROUND AND SIGNIFICANCE

Background

In visual search tasks, observers (Os) look for a target in a visual scene containing distracting stimuli. Some medical screening tasks (e.g. for lung, breast, & cervical cancer) can be described as difficult visual search tasks. A characteristic of these tasks is low *target prevalence.* That is, unlike many other real world search tasks (e.g. finding bananas in the fruit aisle) and unlike typical laboratory search tasks, these are searches for targets that appear only rarely. In routine mammography, for example, pathology is present on less than 1% of breast images. We have found that this low target prevalence, by itself, can be a potent source of miss errors in visual search (Wolfe et al., 2005). In experiments with a range of different search stimuli, we have found that Os miss 0.30 to 0.40 of targets when those targets are present on only 1-2% of search trials. Os miss only 0.05 to 0.10 of the same targets when those targets are present on 50% of search trials.

Visual search tasks can be studied as signal detection problems. Os are trying to distinguish displays containing a target signal from those containing only noise. Miss errors can arise from a lack of sensitivity to the target or from setting a decision criterion to a position that causes detectable targets to be classified as non-targets. In other contexts (e.g. vigilance & categorization literatures), the response to a change in target frequency is understood as a shift in criterion, not in the detectability of the target. We have shown that this is also true in visual search. Low prevalence produces a large shift in criterion that is surprisingly hard to counteract (e.g. by manipulations of costs and benefits for different types of response) (Wolfe et al., 2007).

Because of differences between laboratory search tasks and clinical screening tasks and because clinical tasks are performed by highly trained professionals, it would be unwise to generalize from the existing data to the conclusions about errors in medical screening. Therefore, we have worked to determine whether this is really true for medical screening tasks, and, if so, test theory-based solutions. We have shown prevalence effects in breast cancer (Evans, Birdwell, & Wolfe, 2013) and cervical cancer screening (Evans, Tambouret, Wilbur, Evered, & Wolfe, 2011) as well as in a population of newly-trained airport baggage screeners (Wolfe, Brunelli, Rubinstein, & Horowitz, 2013).

Target-Absent Trials

Understanding low prevalence search requires that we understand behavior on target-absent trials. After all, if the target is rare, then most of the time, the observer will be responding that the target is absent. Understanding these searches for targets that are not present in a display has proven to be one of the enduring problems in the basic and applied search literature. How and when do observers terminate unsuccessful searches? An obvious hypothesis is that observers quit when they have searched through all of the items and determined that none of those items is the target. However, this appealing idea makes predictions that are not supported by the data. For example, we performed experiments in which the items in a search display were randomly replotted every 100 msec during a trial (Horowitz & Wolfe, 1998). In this case, stimuli were letters. Observers were searching for a “T” among “L”s. Our "dynamic search" task makes it impossible to keep track of which distractors have been rejected. Nevertheless, observers were not disabled in their search. Target-absent RTs were slowed relative to a static control condition, but observers had no problem terminating target-absent searches appropriately.

Rather than assuming a sure knowledge that all items have been rejected, a more plausible class of models proposes that observers abandon unsuccessful searches when they conclude that the chance of missing a target is acceptably low. Setting this criterion requires some information about the accuracy of responses. Feedback about accuracy could be given explicitly, but even if it is not, observers will generally know when they have found a target and, it may be presumed, can use this information. Observers must also have some information about the progress of the search. Even if they cannot keep track of the specific items that have been rejected, observers might implicitly count the number items attended or rejected, or they might keep track of the duration of the search. Assuming some sort of meaningful feedback, they could shorten the next search if the prior search was successful or lengthen it if an error was made. Evidence for adaptive criterion setting can be found in sequential RT data. RTs tend to be slightly faster after each successful response and markedly slower after each error (Chun & Wolfe, 1996), just the sort of "staircase" rule one would expect if observers were trying to maintain relatively high accuracy (Levitt, 1971; Rose, Teller, & Rendleman, 1970).

While an adaptive criterion setting method might work quite well in many situations, medical screening tasks pose special challenges. 1) Costs of failure are very high, 2) feedback is delayed or absent (How do you know that you got the ‘right’ answer?), 3) The task is subject to the "Satisfaction of search" problem. This occurs when it is possible to have multiple targets in a single image. Under these conditions, finding one target can lead the observer to abandon the search and, thus, miss other targets in the same display (Samuel, Kundel, Nodine, & Toto, 1995).

Prevalence effects – Previous Experiments and Rationale

As noted, the frequency of target presence is perhaps the most striking difference between typical laboratory search tasks and some of the more important applied tasks. There is a limited literature on the effects of target prevalence in the analysis of medical images. The literature is small for methodological reasons. To quote Gur et al. (2003), "Although there are some indications (and certainly a strong subjective belief among scientists) that there is a “prevalence effect” in the clinic because of (radiologists') expectations that result in certain behaviors, there are hardly any convincing data to demonstrate the effect." A major reason for this dearth is captured by an example described by Kundel (2000). "When the disease prevalence is 0.5, a test set of 100 cases will have a variance approximating .04. Achieving the same variance at a screening prevalence of 0.001 requires a test set of about 25,000 cases!" Achieving enough statistical power in studies with low prevalence requires data sets that do not exist and would be laborious to collect.

The results reported in the small literature that does exist are very mixed, as one might expect given the statistical power issue. In a retrospective examination of lung cancer studies with prevalence ranging from 0.1% to about 50%, Kundel (2000) reports rather large effects on the signal detection measure of detectability, *d'*. The range of *d’* values was 1.4 to 3.4, with higher prevalence yielding lower *d'* values – meaning lower performance. This study involves comparing performance on data collected over many years in many settings. In contrast, Ethell and Manning (2004) did a much smaller study in the lab using wrist fracture X-rays with target prevalence’s of 22%, 50%, & 83% . They reported a modest decrease in A$''$ (area under the ROC, a measure similar to *d’*) as prevalence decreased (.8 at 83%, .71 at 50%, .68 at 22%). Egglin and Feinstein (1996) reported a similar effect with prevalence’s of 20% and 60%. In a larger, well-designed study, Gur et al. (2003) found no significant effect on A$''$ and concluded, "with laboratory conditions, if a prevalence effect exists, it is quite small in magnitude; hence, it will not likely alter conclusions derived from such studies." They were more cautious about generalizing this conclusion to clinical settings (Gur, Rockette, Warfel, et al., 2003). Thus all logical possibilities appear to be represented in the literature and there is no large body of visual search data available to clarify this matter.

Our previous lab data suggests that this may be a dangerous gap in our knowledge. In our first experiments (and, indeed, in most basic search studies in which stimuli remain visible until response), False Alarms are very rare. With those stimuli, under low prevalence conditions, Miss error rates can rise to 30-50%. This rise in miss error rates with decrease in prevalence has been seen before. It is a repeated theme in the extensive work on vigilance in the 1960s. Jane Mackworth (1970) states that, "One of the most important findings in vigilance research has been the discovery that the probability that a signal will be detected is considerably reduced when the background event rate is increased" (see for example Broadbent & Gregory, 1965; Colquhoun, 1961). The "background rate" in this context is the number of non-target events over a period of time in a vigilance task. The more non-target events there are, the lower the target prevalence. The introductory example in Green and Swets' (1967) classic SDT book illustrates that the effect of reducing target prevalence will be to make the decision criterion more conservative. You will be less likely to call something a target if the *a priori* probability of a target is lower.

Note that there is an important difference between traditional vigilance tasks and low prevalence search tasks like mammography. In vigilance, observers are monitoring the world for fleeting stimuli. In mammography and related tasks, each search display with its potential target(s) is present until the observer dismisses it. Nevertheless, we can assume that a similar signal detection analysis can be applied to vigilance tasks and low prevalence search tasks.

In subsequent work, we used stimuli that closely resembled x-rays of baggage such as one might see at the airport. Baggage screening is a difficult task and does produce false alarm errors. This allows us to use signal detection methods in analysis. We find that the prevalence effect is a form of “criterion shift”. Miss errors go up at low prevalence. False alarm errors go down. This is not the pattern of errors one wants to see at the airport or in medical screening. However, as noted above, we have now found prevalence effects in screening for cervical and breast cancer (Evans et al., 2011) (Evans et al., 2013).

The new grant period: Beyond prevalence

In addition to target prevalence, there are other aspects of real-world search tasks, like those in radiology, that can be profitably studied by a combination of basic research in the lab and more translational studies with expert populations. Consider looking at a bone scan or a CT to determine if a patient is responding to cancer treatment. This is a complex visual search task, extending over many minutes, for multiple instances of multiple target types. How do humans perform such tasks and how can we help them to do them better? In the next grant period, we will investigate the basic science of multi-target, extended visual search tasks. The theoretical framework will combine our Guided Search model with aspects of Optimal Foraging Theory and models of working and activated long-term memory. The core idea of Guided Search is that a limited set of fundamental attributes (e.g. color, orientation, motion, etc) guides the deployment of attention (Wolfe, 1994, 2007; Wolfe, Cave, & Franzel, 1989). If you are looking for red vertical lines, attention can be guided to the intersection of the set of red and vertical items. However, such a model, with only a simple target template, is inadequate to a situation where you are searching for multiple instances of several types of target (Find all the big red and curved green items or all the masses, calcifications, and architectural distortions in a mammogram.). We hypothesize that flexible use of working memory in collaboration with activated long-term memory (Oberauer & Hein, 2012) (Cowan, 1995) permits continued guided search and that Foraging Theory (Stephens, Brown, & Ydenberg, 2007; Stephens & Krebs, 1986) will provide an account of when to stop searching. Novel search paradigms will be used to test hypotheses raised by this framework. Moreover, we will translate the basic principles of search to the study of search by experts; specifically, radiologists. We hypothesize that potentially dangerous errors are generated when expert observers are asked to perform artificial search tasks using cognitive processes that evolved to solve very different search tasks. There are three aims, organized around different extended search tasks and the paradigms we will use to study them. For each, we will identify important issues in extended visual search. We will establish the basic science using non-expert observers. Next, we will determine if the rules, established for simple search, apply to extended search. For each aim, we will use expert observers, to test the hypothesis that the default rules of search are a source of error and/or inefficiency in socially important, artificial search tasks like radiology.

There are three new specific aims:

II. SPECIFIC AIMS

Aim1: Foraging: Rather than search for a single target, many radiographic tasks involve search for <u>all</u> instances of a type of target; for example, lung nodules in a chest CT. This type of task is a foraging task, roughly akin to picking berries. There is a substantial literature on animal foraging but much less in humans. We have found that, in a laboratory simulation of berry picking, human behavior is described well by Charnov's (1976) Marginal Value Theorem (MVT). We hypothesize that MVT is deeply engrained in human foraging behavior and that its automatic deployment in medical foraging tasks could be a source of error.

Aim 2: Hybrid search: In many radiographic tasks, Os search for multiple target types, held in a mental checklist. Such searches involve the interaction of visual search with memory search. We call these "Hybrid Searches" We have shown that that hybrid search times increase linearly with increases in the visual set size but log-linearly with increases in the memory set size (Wolfe, 2012). This is true for real-world objects. We hypothesize that this is a general principle of hybrid search. The alternative hypothesis is that real-world objects (and scenes) are special. Thus, other types of memory set items (words, symbols, categories, etc) might not obey this log law. Further, we hypothesize that the pattern of errors is shaped by the relationship between target types leading to a version of a satisfaction of search problem (Berbaum et al., 1990; Berbaum et al., 2013; Cain, Dunsmoor, LaBar, & Mitroff, 2011; Nodine, Krupinski, Kundel, Toto, & Herman, 1992) in which dead ends in memory search yield false negatives in visual search.

Aim 3: Prevalence effects: In the prior grant period, we have shown that changes in the base rate or target prevalence produce changes in criteria that, in turn, produce elevated false negative errors at low prevalence (Wolfe, Horowitz, & Kenner, 2005; Wolfe et al., 2007; Wolfe & VanWert, 2010). We have shown that these prevalence effects occur when expert searchers perform in their domain of expertise (Evans et al., 2011; Wolfe, Birdwell, & Evans, 2011). In the next grant period, we will test proposed interventions to reduce these errors. In addition, we will examine the interaction of prevalence with the foraging and hybrid search conditions of Aims 1 and 2.

(Bond, 1981; Burger, Huang, Hemerik, van Lenteren, & Vet, 2006; Cain, Clark, Mitroff, & Vul, 2011; e.g. Charnov, 1976; Kundel & Nodine, 1975; Nodine & Mello-Thoms, 2010; e.g. Pyke, Pulliam, & Charnov, 1977; but see Smith, Hood, & Gilchrist, 2008; Wolfe & VanWert, 2010)

Additional Studies

Our work on prevalence effects has led into studying other aspects of mammography. Specifically, we are measuring the eye movements of radiologists as they examine breast tomosynthesis images (methods described below). We are also studying two different ways of comparing this year's images with a prior image in breast imaging. The two images can be viewed side-by-side or they can be alternated ("toggled") in the same location, aligned as best possible. We are testing the hypothesis that the later 'toggling' technique will be faster and/or more accurate than the side-by-side approach. In both of these studies, one set of experiments is run on non-radiologists in the lab. Other studies, with radiologist observers are conducted either in our facility in BWH or at national meetings of radiologists (e.g. RSNA, ECR). Radiologists may be asked to evaluate medical images on various workstations (e.g., Automated Breast Ultrasound, or ABUS) while we track their eye movements, or monitor for keypresses and mouse clicks.

III. SUBJECT SELECTION

Ten to 15 subjects per study are typical in experiments of this sort, and reviewers of our papers and grants will not be satisfied with fewer. In addition, data from about 33% of our subjects is discarded regularly because subjects did not complete the task or simply did not do the correct task. Each subject must report no history of eye or muscle disorders, as well as 20/25 vision or better, and pass the Ishihara Tests for color-blindness. The subjects who participate in online studies (through services such as Mechanical Turk) will not be tested for colorblindness, and will not have their visual acuity checked. They will attest to having 20/25 vision or better and have "normal" color vision in order to participate. For online experiments, we try to build in and pre-register quality checks that allow us to exclude participants after they have been test, even if they cannot be screened before participation. Note that there is no risk to the participant if they don't meet our criteria but are still tested. The risk is to data quality.

Subjects are recruited with posters placed in local colleges and in the lab neighborhood, as well as with ads placed on public access websites such as Craigslist. Subjects are also recruited from the Harvard Decision Sciences Lab participant pool. Subjects from this pool voluntarily subscribe to an email list and are alerted to upcoming studies. It is anticipated that equal numbers of women and men will be included in this study. Our recruitment procedures are unbiased and consider all applicants regardless of race, color, creed, or national origin.

For online studies through services like Mechanical Turk, most participants will sign up for these studies as they see them online through websites and will choose to participate in our experiments of their own volition and interest. In some cases (e.g. when trying to recruit special populations like pathologists, we may advertise via appropriate social media or email lists).

We also recruit experts including medical technicians (e.g. cytologists) and MDs (typically radiologists). As a general rule, these are recruited by one-to-one conversations at Brigham and Women's hospital as well as other local hospitals (e.g. Tufts). Radiologists at the UT MD Anderson Cancer Center were recruited via emails sent by the head of the radiology department on our behalf. We also have some experiments that are run as exhibits or educational displays at national scientific meetings (E.g. the Radiological Society of North America – RSNA – annual meeting in Chicago, European Congress of Radiology – ECR – annual meeting in Vienna, Austria). Here our observers are volunteers who happen to be attracted to our presentation.

IV. SUBJECT ENROLLMENT

Upon passing selection criteria (above), subjects read a consent form, and verbally agree to it. Subjects are told about the general nature of the experiment prior to testing and a more detailed explanation of the purposes of the study after enrollment. General subjects for in-person studies are compensated with $15/hour for participation. In online experiments, subjects will read a consent fact sheet for online study. They will also be provided with information about the study before they check boxes stating that they are over 18 years old, meet the inclusion requirements, read through the consent sheet and consent to participate. Subjects for on-line studies are typically compensated with $8/hour for participation though some online studies are framed as contests where participants are informed in advance of possible prizes but not otherwise compensated. When the enrollment of specialists, such as cytotechnologists or mammography screeners is required, the participants are recruited from Partners hospitals and other, local biomedical facilities. The compensation for technical staff is $35/hour, because they are trained experts rather than naïve observers. The expert screeners who participate during their normal workflow are not given financial reimbursement for their participation. Radiologists are sometimes compensated with good chocolate or some similar gift. At the experiments run at RSNA, ECR or other radiology meetings, radiologists may be entered into a contest where the best score is rewarded with an iPad.

V. STUDY PROCEDURES

In these experiments, visual stimuli (e.g., objects, scenes, simulated x-rays, magnified cell culture slide photographs, breast x-rays) are presented on a computer screen. On each trial, subjects examine the stimuli and indicate as quickly and accurately whether a target (e.g., a "tumor" or a "wrench") is present or absent by pressing designated buttons. In some studies, observers will localize or describe targets rather than simply declaring them to be present or absent. In some experiments, there will be many different types of target (e.g. look for a chicken, a screw, a door, or a cup). In some experiments, there will be multiple instances of targets in the same display (e.g. find all the lung nodules in this chest x-ray). Search stimuli are usually continuously visible until observers choose to respond. In some experiments, stimuli are only briefly presented. Observers may be given feedback on their performance. This could include a point system designed to emphasize the importance of finding the target (e.g., negative points for false alarms and misses, as well as subtraction of 1 point/second, and positive points for hits). It could include a monetary payoff with money added for correct response and subtracted for errors. In no case would participant payment fall below $8/hr. Subjects are typically given 50-100 practice trials and then tested in blocks of 200 - 1000 experimental trials with breaks in between and within blocks. Subjects are instructed before the task and debriefed after completion of the experiment. Some of our studies might also use a harmless (see below) eyetracker device, which measures eye movements. If eye movements are monitored when subjects are doing the task, an instrument will be recording the reflection of a beam of near-infrared light from their eye. This beam gives information about where their eyes are looking. Some of our studies may be conducted on-line. The experiment procedures will be very similar to those conducted in-person; however, eye movements will not be monitored in these studies. We will record observers' responses and reaction times. The experiments are hosted on-line, and subjects can access the experiment using a web link. Sometimes, an experimenter may assist or monitor the subjects during the study via an on-line video platform, e.g. Zoom. Online studies will provide assurances that we will not collect personal information or use the information for anything other than our scientific purposes. The initial information will disclose the use of ‘cookies’ to determine if the participant has performed the task before. Potential participants will have unlimited time to decide and will not be able to continue using the web site without their consent. To begin the experiment, participants will need to check boxes stating that they read through the consent sheet and agree to participate. Completion of this online form will be taken as consent to participate.

VI. BIOSTATISTICAL ANALYSIS

Error rates and response times (RTs) are the primary measures in these experiments. The nature of the stimuli, the number of items presented (the “set size”) and the frequency of target presence (the “target prevalence”) are varied across trials and blocks, respectively. Typically, trials are discarded if the RT is less than 200 ms and greater than 10000 ms. Interactions of interest include RT and prevalence, error rate and prevalence, target presence/absence and error rate, set size and error rate, set size and RT, and target type and error rate. All data are usually analyzed for statistical significance using analysis of variance (ANOVA) methods. We also use *d'* and criterion, the signal detection measures of detectability and bias to analyze performance. In experiments using the eyetracker, we are usually interested in number of fixations, fixation duration, and scan path.

VII. RISKS AND DISCOMFORTS

The only stresses and aggravations of these experiments are psychological. Some subjects find the tasks boring, while others find them challenging. Experts may have a particular, competitive desire to get the “right” answer. Occasionally, the sustained attention required can cause some subjects to have headaches or feel sleepy. Subjects are free to pause at any time, and may withdraw from the experiments if they wish.

In some of the experiments we will measure eye movements. There are no known risks associated with the eye movement system (pupil centre to corneal reflection system) to be used in these experiments. Illumination of the eye in these systems is less than 1 milliwat per cm squared of near infrared light. Federal safety standards indicate that levels of 10 milliwatts per cm squared are not hazardous to adults or children.

VIII. POTENTIAL BENEFITS

There are no specific health benefits for participation in these experiments though subjects may learn about their visual system and experts may learn about behaviors that are important to their work. This research serves to benefit society by increasing our understanding of socially important search contexts such as airport security and medical screening.

IX. MONITORING AND QUALITY ASSURANCE

Subjects are free to pause the experiment or withdraw from participation at any time. Serious and unexpected adverse events will be reported immediately (within 24 hours) to the Human Research Committee by telephone, fax or e-mail and will be followed by a complete written report within 10 working days.

The privacy of subjects will be protected and their results will remain confidential. To assure this, subject identifiers are removed from the data files. Ultimate responsibility for data quality rests with the Principal Investigator, Jeremy Wolfe. Research assistants monitor the data as it is collected to assure its integrity.

Some of our experiments will employ modified medical screening materials, such as breast x-rays or photographs of magnified cell slides. All of these materials will be anonymized prior to being used in any of the experiments by experts in the BWH radiology department. No patient data whatsoever will be associated with the images.

X. REFERENCES

Berbaum, K. S., Franken, E. A., Jr., Dorfman, D. D., Rooholamini, S. A., Kathol, M. H., Barloon, T. J., et al. (1990). Satisfaction of search in diagnostic radiology. *Invest Radiol, 25*(2), 133-140.

Berbaum, K. S., Schartz, K. M., Caldwell, R. T., Madsen, M. T., Thompson, B. H., Mullan, B. F., et al. (2013). Satisfaction of Search from Detection of Pulmonary Nodules in Computed Tomography of the Chest. *Acad Radiol, in press*.

Bond, A. B. (1981). Giving-up as a Poisson process: the departure decision of the green lacewing. . *Animal Behaviour, 29*, 629-630.

Broadbent, D. E., & Gregory, M. (1965). Effects of noise and of signal rate upon vigilance analysed by means of decision theory. *Hum Factors, 7*(2), 155-162.

Burger, J. M. S., Huang, Y., Hemerik, L., van Lenteren, J. C., & Vet, L. E. M. (2006). Flexible use of patch-leaving mechanisms in a parasitoid wasp. *Journal of Insect Behavior, 19*(2), 155-170.

Cain, M. S., Clark, K., Mitroff, S. R., & Vul, E. (2011). *An Optimal Foraging Model of Human Visual Search.* Paper presented at the Cognitive Science Society, Boston.

Cain, M. S., Dunsmoor, J. E., LaBar, K. S., & Mitroff, S. R. (2011). Anticipatory Anxiety Hinders Detection of a Second Target in Dual-Target Search. *Psychological Science, 22*(7), 866–871.

Charnov, E. L. (1976). Optimal foraging, the marginal value theorem. . *Theoretical Population Biology, 9*, 129-136.

Chun, M. M., & Wolfe, J. M. (1996). Just say no: How are visual searches terminated when there is no target present? *Cognitive Psychology, 30*, 39-78.

Colquhoun, W. P. (1961). The effect of 'unwanted' signals on performance in a vigilance task. *Ergonomics, 4*, 41-51.

Cowan, N. (1995). *Attention and Memory: An integrated framework.* New York: Oxford U press.

Egglin, T. K., & Feinstein, A. R. (1996). Context bias. A problem in diagnostic radiology. *Jama, 276*(21), 1752-1755.

Evans, K. K., Birdwell, R. L., & Wolfe, J. M. (2013). If You Don't Find It Often, You Often Don't Find It: Why Some Cancers Are Missed in Breast Cancer Screening. . *PLoS ONE 8(5): e64366. , 8*(5), e64366.

Evans, K. K., Tambouret, R., Wilbur, D. C., Evered, A., & Wolfe, J. M. (2011). Prevalence of Abnormalities Influences Cytologists' Error Rates in Screening for Cervical Cancer" *Archives of Pathology & Laboratory Medicine, 135*(12), 1557-1560. .

Green, D. M., & Swets, J. A. (1967). *Signal detection theory and psychophysics*. New York: John Wiley and Sons.

Gur, D., Rockette, H. E., Armfield, D. R., Blachar, A., Bogan, J. K., Brancatelli, G., et al. (2003). Prevalence effect in a laboratory environment. *Radiology, 228*(1), 10-14.

Gur, D., Rockette, H. E., Warfel, T., Lacomis, J. M., & Fuhrman, C. R. (2003). From the laboratory to the clinic: the "prevalence effect". *Acad Radiol, 10*(11), 1324-1326.

Horowitz, T. S., & Wolfe, J. M. (1998). Visual search has no memory. *Nature, 394*(Aug 6), 575-577.

Kundel, H. L. (2000). Disease prevalence and the index of detectability: a survey of studies of lung cancer detection by chest radiography. In E. A. Krupinski (Ed.), *Medical Imaging 2000: Image Perception and Performance* (Vol. 3981, pp. 135-144).

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

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *J. Acoust. Soc. Am., 49*, 467-477.

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

Nodine, C. F., Krupinski, E. A., Kundel, H. L., Toto, L., & Herman, G. T. (1992). Satisfaction of search (SOS). *Invest Radiol, 27*(7), 571-573.

Nodine, C. F., & Mello-Thoms, C. (2010). The role of expertise in radiologic image interpretation. In E. A. Krupinski & E. Samei (Eds.), *The Handbook of Medical Image Perception and Techniques* (pp. 139-156). Cambridge, UK: Cambridge U Press.

Oberauer, K., & Hein, L. (2012). Attention to information in working memory. *Current Directions in Psychological Science, 21*(3), 164-169.

Pyke, G. H., Pulliam, H. R., & Charnov, E. L. (1977). Optimal Foraging: A Selective Review of Theory and Tests. *The Quarterly Review of Biology, 52*(2), 137-154.

Rose, R. M., Teller, D. Y., & Rendleman, P. (1970). Statistical properties of staircase estimates. *Perception and Psychophysics, 8*(4), 199-204.

Samuel, S., Kundel, H. L., Nodine, C. F., & Toto, L. C. (1995). Mechanism of satisfaction of search: eye position recordings in the reading of chest radiographs. *Radiology, 194*(3), 895-902.

Smith, A. D., Hood, B. M., & Gilchrist, I. D. (2008). Visual search and foraging compared in a large-scale search task *Cognitive Processing* (Vol. 9, pp. 121-126): Springer Heidelberg.

Stephens, D. W., Brown, J. S., & Ydenberg, R. C. (2007). *Foraging: Behavior and Ecology*. Chicago: U. Chicago Press.

Stephens, D. W., & Krebs, J. R. (1986). *Foraging Theory*. Princeton, NJ: Princeton U. Press.

Wolfe, J. M. (1994). Guided Search 2.0: A revised model of visual search. *Psychonomic Bulletin and Review, 1*(2), 202-238.

Wolfe, J. M. (2007). Guided Search 4.0: Current Progress with a model of visual search. In W. Gray (Ed.), *Integrated Models of Cognitive Systems* (pp. 99-119). New York: Oxford.

Wolfe, J. M. (2012). Saved by a log: How do humans perform hybrid visual and memory search? *Psychol Sci, 23*(7), 698-703.

Wolfe, J. M., Birdwell, R. L., & Evans, K. K. (2011). If You Don't Find it Often, You Often Don't Find It: Disease Prevalence is a Source of Miss Errors in Screening Mammography. *paper presented at the Anuual Radiological Society of North America meeting: Chicago, IL., Nov 30, 2011*.

Wolfe, J. M., Brunelli, D. N., Rubinstein, J., & Horowitz, T. S. (2013). Prevalence effects in newly trained airport checkpoint screeners: Trained observers miss rare targets, too. *Journal of Vision, 13*(3:33), 1-9.

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided Search: An alternative to the Feature Integration model for visual search. *J. Exp. Psychol. - Human Perception and Perf., 15*, 419-433.

Wolfe, J. M., Horowitz, T. S., & Kenner, N. M. (2005). Rare targets are often missed in visual search. *Nature, 435*(7041), 439-440.

Wolfe, J. M., Horowitz , T. S., VanWert, M. J., Kenner, N. M., Place, S. S., & Kibbi, N. (2007). Low target prevalence is a stubborn source of errors in visual search tasks. *JEP: General, 136*(4), 623-638.

Wolfe, J. M., & VanWert, M. J. (2010). Varying target prevalence reveals two, dissociable decision criteria in visual search. . *Curr Biol, 20*(2), 121-124.

SPECIFIC PROTOCOL

Derived from: Wolfe, J. M. (2022). How one block of trials influences the next: Persistent effects of disease prevalence and feedback on decisions about images of skin lesions in a large online study. *Cognitive Research: Principles and Implications (CRPI), 7*, 10. doi: https://doi.org/10.1186/s41235-022-00362-0

In visual decisions about finding and/or identifying a target, the prevalence of the target makes a difference (Horowitz, 2017). By “prevalence”, we mean the frequency with which a target appears in a series of trials or cases. The effects of prevalence are of more than academic interest because target prevalence can vary dramatically across tasks in the real world. For example, in a task like identifying signs of breast cancer in mammographic images, the prevalence is very low in a breast cancer screening program where cancer might be present on 0.5% of images and where findings that are suspicious enough to require more testing might be present on 5-10% of cases (e.g. Jackson, et al., 2012). The prevalence of a disease will be much higher in a set of images referred to the radiologist because an initial screening was suspicious. The classic low prevalence effect (LPE) involves an increase in false negative / miss errors and, usually, a decrease in the rate of false positive / false alarm errors (Wolfe, Horowitz, & Kenner, 2005; Wolfe, et al., 2007). In signal detection terms, the LPE can be described largely, but not entirely, as a “criterion shift” in which participants become more ‘conservative’ about declaring a target to be present (Hautus, Macmillan, and Creelman, 2021) . In visual search studies of the LPE, participants also tend to abandon search more quickly when targets are not found (Wolfe & Van Wert, 2010).

We also know that what you have seen influences what you will report seeing next, for example, in classic adaptation effects (e.g. Helson, 1964) or in serial dependence effects like those studied by Fischer & Whitney (2014) and many others (e.g. Gekas, McDermott, & Mamassian, 2019 or Manassi, et al., 2021 for work with radiologists). What you are *told* that you will see also has an impact. Thus, you will look harder if you are given information that suggests that it is like that there is something to find (Reed, Chow, Chew, & Brennan, 2014; Littlefair, et al., 2016). In this paper, we examine these effects together. What is the effect of the prevalence in one block on performance on a subsequent block?

Our understanding of prevalence effects was complicated in 2018, when Levari et al. (2018) reported that it was possible to obtain effects in the opposite direction from the classic LPE. Their participants made decisions about a single item on a continuum. For instance, participants might be asked if a dot, drawn from a blue-purple color continuum, was 'blue'. When fewer dots were drawn from the blue end of the continuum, they reported that participants became more liberal about calling ambiguous dots 'blue'. They called this effect "Prevalence-induced concept change" (PICC). Lyu et al. (2021) found that one driver of these opposing LPE and PICC results was the presence or absence of feedback. When making decisions about the same stimulus continuum, participants reliably produced LPE effects when given feedback after each trial. They tended to produce PICC effects (somewhat less reliably) when there was no feedback. This has real-world implications because, just as tasks differ in target prevalence, they differ in feedback. For instance, in training, participants might receive immediate feedback after every trial. In the field, that feedback might be delayed, partial, or unavailable. Consider airport security screening. In training, participants are likely to see targets at relatively high prevalence, with feedback. At the airport checkpoint, real 'threats' will be rare (we may hope!) and security screens may get some feedback about some positive cases because the suspect bag is opened on the spot. False negative errors probably generate no feedback, even though these would be the most serious errors in a security setting.

The central question of the present study is how experience with one level of prevalence, with or without feedback, influences performance on subsequent trials where either prevalence and/or feedback conditions could have changed. In the great bulk of research on prevalence effects, high prevalence trials were followed by low prevalence trials or participants experienced only a single prevalence level. Typically, feedback is not independently manipulated (but see Growns & Kukucka, 2021; Lyu, et al., 2021; Papesh, Heisick, & Warner, 2018; Weatherford, Erickson, Thomas, Walker, & Schein, 2020). Moreover, in previous studies, the change due to a previous discrete block of trials was not assessed.

In the present study, prevalence can be high or low and feedback can be present or absent. This yields four conditions: Low prevalence with feedback, Low prevalence without feedback, High prevalence with feedback, and High prevalence without feedback. There are 16 possible pairings of two consecutive blocks. In this study, the stimuli are skin lesions: either nevi (skin 'moles'; singular, nevus) or melanoma (skin cancer). A large dataset of over 300,000 decisions was collected on-line by participants using a medical image labeling app (“DiagnosUs” http://diagnosus.com/). Participants had varying levels of expertise from complete novice to MD as will be described later. Thus, in addition to allowing us to assess the influence of one prevalence X feedback combination on another, these data provide new evidence about prevalence effects in expert populations (K K Evans, Birdwell, & Wolfe, 2013; K. K. Evans, Tambouret, Wilbur, Evered, & Wolfe, 2011; Evered, 2017; Mitroff, Ericson, & Sharpe, 2017;

Reed, Ryan, McEntee, Evanoff, & Brennan, 2011; Trueblood, et al., 2021; Wolfe, Brunelli, Rubinstein, & Horowitz 2013).

To anticipate the broad outlines of the results, LPE and weak PICC effects are produced with these dermatology stimuli. As in other studies with expert populations, our experts show these effects, as do novices. In terms of the influence of one block of trials on the next block, we find that, in general, experience with a low prevalence block with feedback makes one more conservative in any subsequent block while the experience of high prevalence with feedback makes one more liberal. A block without feedback does not appear to have a significant impact on the next block.

Methods

Data were collected on-line via a free and open iOS application, "DiagnosUs", created by Centaur Labs as a platform for game-like image labeling competitions https://www.centaurlabs.com/). Between 6/22/21 and 6/27/21, Centaur Labs ran 24 'contests' on our behalf, collecting >300,000 trials from 803 participants in 6 days. Each of the 24 contests consisted of 80 unique images of a skin lesion as shown in Figure 1. This is a relatively small number of images, but it kept the sessions short and encouraged participation. Images came from the International Skin Imaging Collaboration (ISIC) 2018 challenge (Codella, et al., 2019). The set contains over 1,200 melanoma images and over 7,600 nevus images.

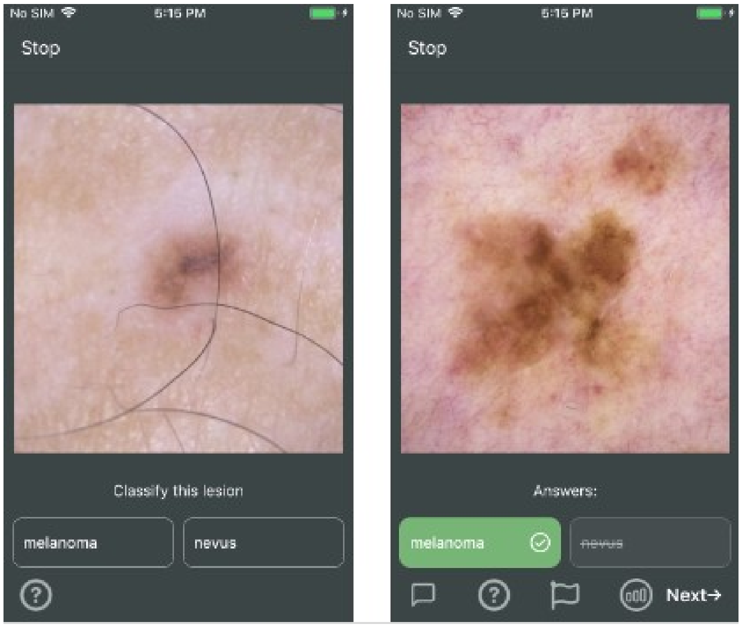


Figure 1: Sample stimulus displays. Left: Participants are asked if a lesion is a melanoma (cancer) or a nevus (benign 'mole'). Right: After response, if this were a feedback condition, red or green feedback would inform the observer of the correctness of the answer. In no feedback conditions, neutral feedback indicated that the response had been registered.

On each of the six days, there were four contests available. These were:

1) Low Prevalence – No Feedback; consisting of 20% (16) melanoma images
2) Low Prevalence – Feedback; prevalence was the same as in #1, but with accurate, trial by trial feedback
3) High Prevalence – No Feedback; consisting of 50% (40) melanoma images
4) High Prevalence – Feedback; prevalence was the same as in #3, but with accurate, trial by trial feedback

The contests were randomly assigned to participants with the constraint that they did not run the same type of contest more than once in a day. Participants could take part in as many or as few contests as they wished over the course of the six days and they could end up repeating a condition from one day on another day. The prevalence and feedback conditions were not advertised to participants. Accuracy is the main dependent variable though we also collected response times.

Because of the on-line, voluntary nature of the data collection, we did not have control over the viewing conditions or the type of screen used. We compensated by collecting a very large dataset (see below).

Observers

We collected data from 803 unique individuals. For each individual, we have a crude categorization of expertise as shown in Table 1:

Medical Student	337
Pre-Med Student	174
No Medical Experience	145
Medical Doctor	54
Other Healthcare Professional	47
Nurse	25
Medical Technician	11
Physician Assistant	10

Table 1: Observers divided by expertise category

Participants were asked why they chose to participate and given three choices, as shown in Table 2.

Compete With Others	37
Earn Money	248
Improve My Skills	518

Table 2: Observers divided by reason to participate

Participants reported coming from 78(!) countries with three quarters of the participants coming from nine countries, as shown in Table 3.

Philippines	198
USA	182
Ghana	94
Great Britain	31
Romania	30
Canada	27
Mexico	19
Indonesia	15
Australia	13
All other	194

Table 3: Reported country of origin.

Beyond this demographic information, we have only an observer number. We were given no identifying information about the participants. Participation on the app constituted consent. Procedures were approved by the Institutional Review Board at Brigham and Women's Hospital (IRB #2007P000646).

These 803 participants produced 311,842 trials of data. An excel spreadsheet with all data is posted on the Open Science Framework at https://osf.io/hck5n/. From this set, we eliminated participants who did not complete a full block of 80 trials. This left 630 participants who ran a total of 277,371 trials in 2,988 blocks. Different participants chose to participate in different numbers of blocks. As shown in Table 4.

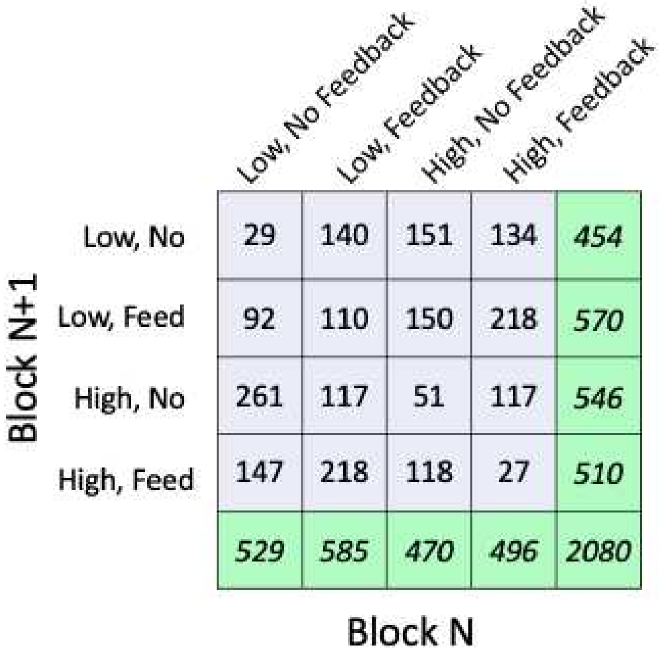
Blocks Run	# of Participants
1	315
2	178
3	104
4	56
5	52
6	22
7	27
8	20
9	14
10	10
11	10
12	5
13	10
14	2
15	4
16	6
17	5
18	3
19	10
20	3
21	4
22	2
23	6
24	4

Table 4: Numbers of participants who ran N (1-24) blocks over 6 days of testing.

Those participants running only a single block were removed from most analyses because our primary interest is the influence of one block upon the next. Obviously, with some participants contributing 24 blocks and others contributing 2, the dataset is unbalanced. We repeated the main analyses reported below, limiting the analysis to only the first pair of blocks for each observer. This reduces power. However, the main patterns of results, reported below, are found when the dataset is restricted to only blocks 1 and 2. Accordingly, we think that the unequal numbers of trials and blocks from different participants is not a significant issue for the present study. In the results reported, we analyze results from the 2,998 remaining blocks of data.

Statistical Analysis / Results

To analyze the effect of one block on the next block, we derived all the pairs of blocks in the dataset. For each of the blocks, we calculated the true positive and false positive proportions and used these to derive the standard signal detection measure of d’ and criterion (Note that we are refraining from using the term “sensitivity” because it is used to refer to d’ in the behavioral science community and to the true positive rate in the medical community.) For the primary analyses, pairs were removed from analysis if either member of the pair produced a d’ of less than 0.5. The task was relatively difficult and we had little control over the motivation of participants in online study of this sort. The probability of achieving a d' > 0.5 by guessing through an 80 trial block is ~0.6%. A cutoff of 0.4 would permit over 2% of blocks to be pure guessing. A d'> 0.5 cutoff left 2,080 pairs of blocks. The distribution of those pairs is shown in Table 5:



Block N+1 \ Block N	Low, No Feedback	Low, Feedback	High, No Feedback	High, Feedback	
Low, No	29	140	151	134	*454*
Low, Feed	92	110	150	218	*570*
High, No	261	117	51	117	*546*
High, Feed	147	218	118	27	*510*
	529	*585*	*470*	*496*	*2080*

Table 5: Counts of the 16 different pairs of Block N and Block N+1.

As can be seen, there is a good, if uneven distribution of pairs. There are fewer pairs on the diagonal where the condition is the same in Blocks N and N+1. In order to run the same block, twice in a row, participants needed to encounter that condition last on Day K and first on Day K+1 since the same condition could not be run twice on the same day. The marginals show that this distribution of the four different conditions is very roughly uniform.

First, we compare performance in the four conditions, regardless of their position the sequence of sessions for each observer. Figure 2 shows the signal detection measures of d' and criterion, "c". These values cannot be computed if p(True Positive) = 1.0 or p(False Positive) = 0.0. Accordingly, in keeping with one standard practice, we add ½ of a False Positive error to each False Positive count (Hautus, Macmillan, & Creelman, 2021).

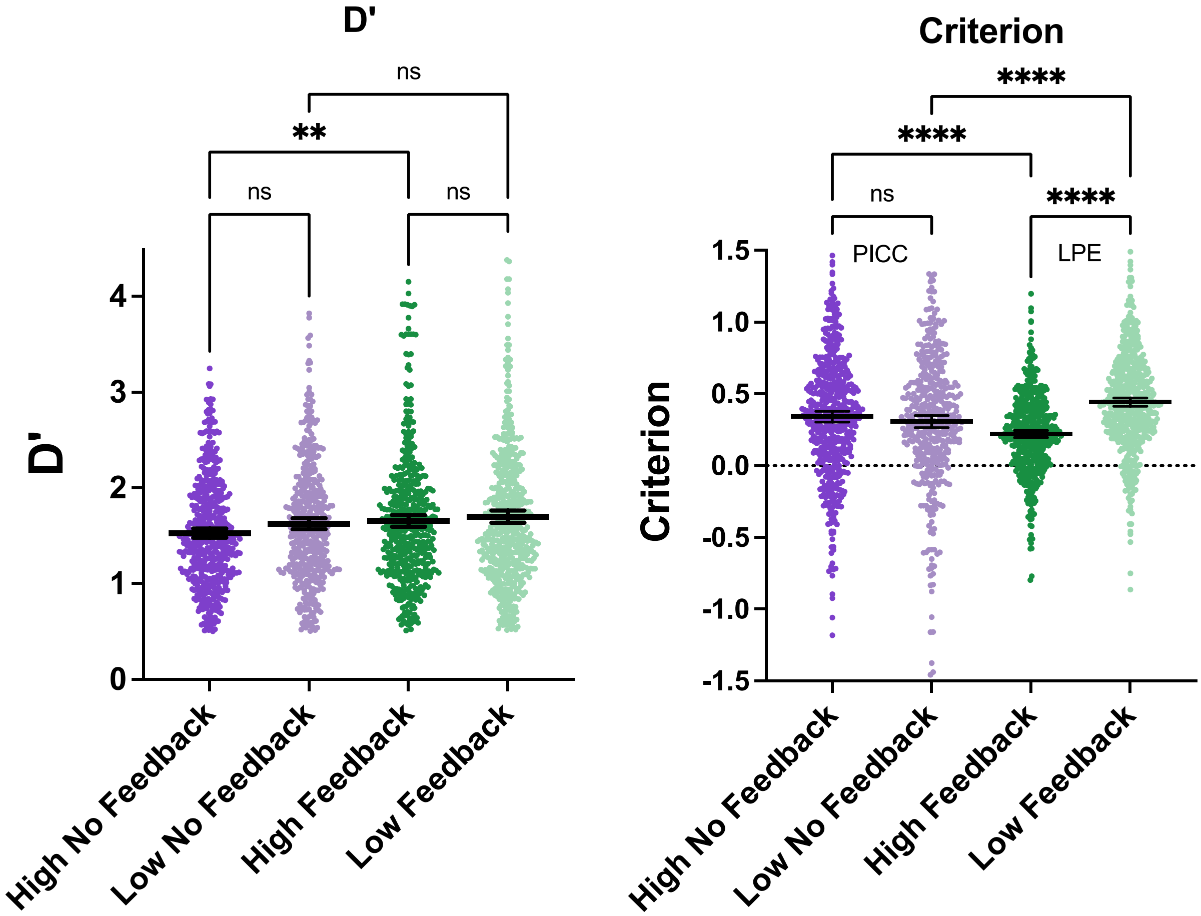


Figure 2: D' and Criterion as a function of condition. Each dot represents one of 2080 blocks of data. Black lines show mean and +/- 95% CI of the mean.

A one-way ANOVA shows a main effect of condition on d' (F(3, 2076) = 6.461, p = 0.0002, partial eta-sq = 0.01) and Šídák's multiple comparisons test shows that performance on skin cancer detection task is modestly better in the high prevalence, feedback condition than in the high prevalence, no-feedback condition (p = 0.008). Prevalence effects are better understood as criterion shifts. An ANOVA shows a very large main effect of condition on criterion (F(3, 2076) = 30.16, p < 0.0001, partial eta-sq = 0.04). Pairwise comparisons show a strong LPE effect when feedback is given (p < 0.0001). Criterion becomes more conservative at low prevalence. Without feedback, criterion becomes slightly more liberal on average (0.34 to 0.31). This is not statistically significant (p = 0.47). As noted earlier, the PICC effect (Levari, et al., 2018), while certainly real, is typically more fragile than the LPE effect in the opposite direction (Lyu, et al., 2021). The differences between conditions with feedback and without feedback are highly significant (p < 0.0001 for each comparison). A significant LPE effect is seen if analysis is restricted to each participant's first block of data (p < 0.0001). There is no PICC effect. If all blocks of data are used, eliminating the d' > 0.5 filter, there is, again a significant LPE (p < 0.0001) and an insignificant PICC effect (p = 0.1244).

Even though participants were not focused on the speed of their responses, response time data reflect the effects of prevalence and feedback. Figure 3A shows the distribution of the ~300,000 trials, divided by block type. As is typical with RT distributions, these distributions have very long positive tails. Some very long RTs probably reflect extended thought, while others might reflect checking an incoming text message. We chose to eliminate the longest 1% of RTs as outliers (>7 sec). Figure 3B shows the mean and 95% confidence intervals on a much finer scale so that the differences between conditions are more clearly visible. There is a significant effect of block type (F(3, 299422) = 52.88, p < 0.0001, partial eta-sq = 0. 0.0005). Šídák's multiple comparisons test shows that the 46 msec difference between high and low prevalence with feedback is significant (p < 0.0001). Low prevalence RTs are shorter than high prevalence. Without feedback, the PICC effect on RT goes in the opposite direction. Low prevalence RTs are 18 msec longer than high prevalence (p = 0.0008). These effects are in line with the RT effects seen in other prevalence studies (Wolfe & VanWert, 2010).

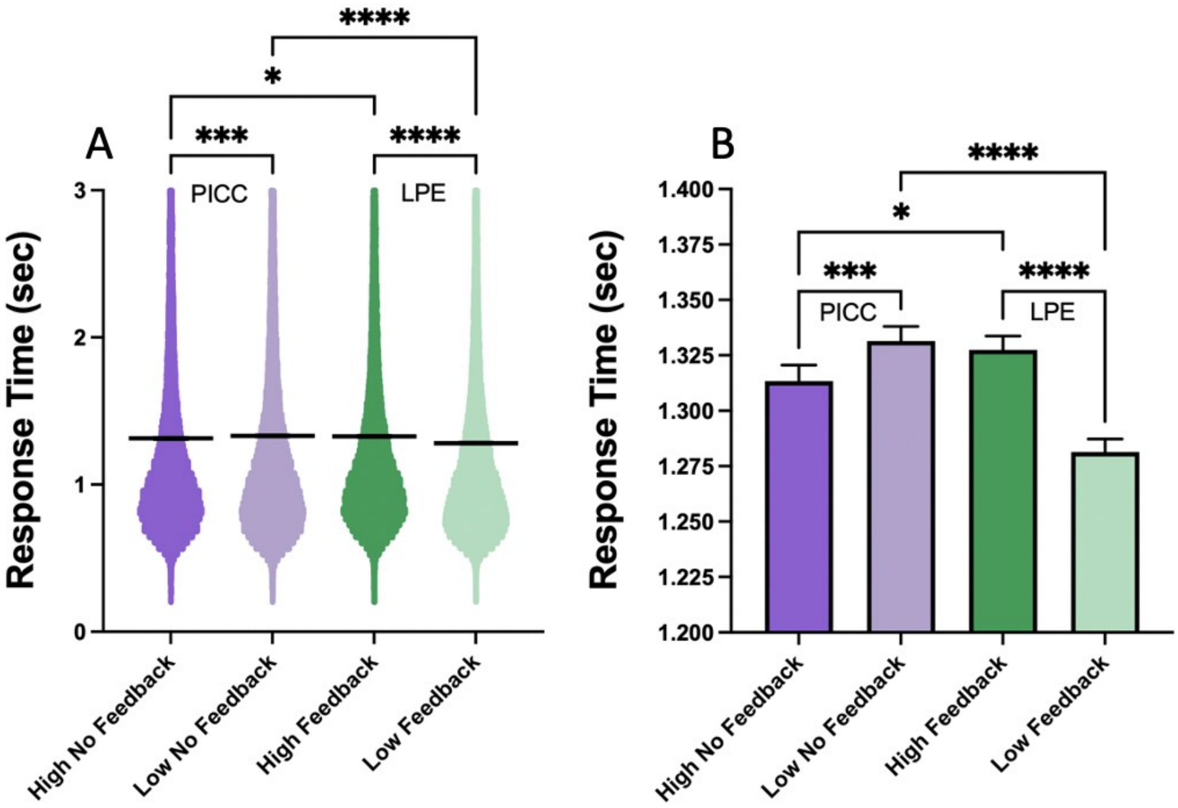


Figure 3: Response time as a function of block type. RTs < 7 sec were analyzed. (A) RT distributions with the graph truncated at 3 sec. Black lines show the mean. (B) Mean RT on a much finer scale. Error bars show +95% confidence intervals.

How does one block influence the next?

The block data essentially replicate previous results, collected under very different conditions. However, the key question for this study is whether or not exposure to one combination of prevalence and feedback influences the next block of trials. Blocks could be independent of each other. The effect of one block might have a transient impact on the next block (e.g. lasting for a few trials and then fading). Finally, the effect of one block on the next could be dependent on the delay between blocks (especially since that delay could be from one day to the next.). To assess each of these effects, we derive measures of d’ and criterion for each of the 16 pairs (combinations of one of four ‘first blocks’ with one of four ‘second blocks’). We then subtract a baseline derived from all blocks of the second block condition. We use the second blocks as the baseline because we are looking for a change in the second block. In fact, it doesn’t make much difference. Many second blocks in one pair are the first blocks in a subsequent pair and the baselines would be very nearly the same if all blocks were included as in Figure 2 above. Thus, for example, to assess the influence on criterion of Low Prevalence without Feedback (Condition 1 in the figures to follow) on High Prevalence without Feedback (Cond 3), we compute Criterion(pair 13) – Criterion(Cond 3). This is done for each of the 2080 pairs of blocks using the d' > 0.5 filter. The data for each of these 2080 "good" pairs of blocks are posted on the Open Science Framework at https://osf.io/hck5n/.

Figure 4 shows the change in D’ as a function of pair with each dot representing one pair of blocks. Statistical tests are simple T-tests against a null hypothesis of no change in d’ from block 1 to block 2. Three pairs, forming no obvious pattern, reach statistical significance (all p = ~0.02, all Cohen's d = ~0.2). As these are not corrected for multiple comparisons, these should not be considered strong effects. The general picture is of little or no effect of the first block of trials on d’ in the second block of trials.

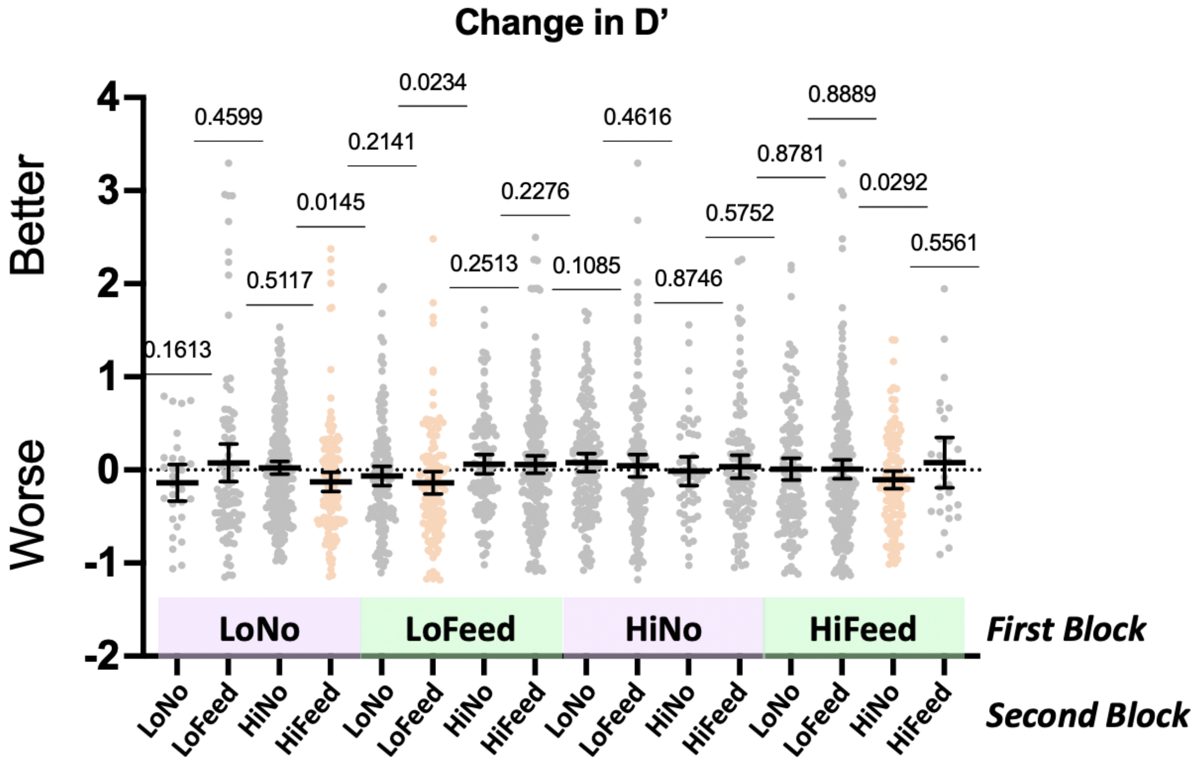


Figure 4: Change in d' on block 2 as a function of the nature of block 1. X-axis gives numerical values for each member of a pair where 1 = low prevalence, no feedback, 2 = low prevalence, with feedback, 3 = high prevalence, no feedback, and 4 = high prevalence, with feedback. P-values show results for simple t-tests, testing against the null hypothesis that there is no change in d'. Black lines show means +/- 95% confidence intervals.

A different pattern is seen in the criterion data, plotted in Figure 5. Here, there are two clear clusters of significant effects. When block 1 has low prevalence and feedback, criterion on block 2 is more conservative (Pairs where Block 1 is low prevalence with feedback, all $p < 0.006$, Cohen's d between 0.2 and 0.5). When block 1 has high prevalence and feedback, criterion on block 2 is more liberal (Pairs where Block 1 is high prevalence with feedback, $p < 0.0005$, pair

44, p = 0.017; note that there are only 27 pairs of this variety in the dataset, All Cohen's d between 0.2 and 0.5). When block 1 does not have feedback, there is no effect on criterion in pair 2, except, perhaps, for 'high feedback' on Block 2 (p < 0.028, again, not corrected for multiple comparisons, Cohen's d = 0.18). This is the main finding of the paper. Criterion can be manipulated by the prevalence of the preceding block of trials, but only if the participants received feedback on that first block.

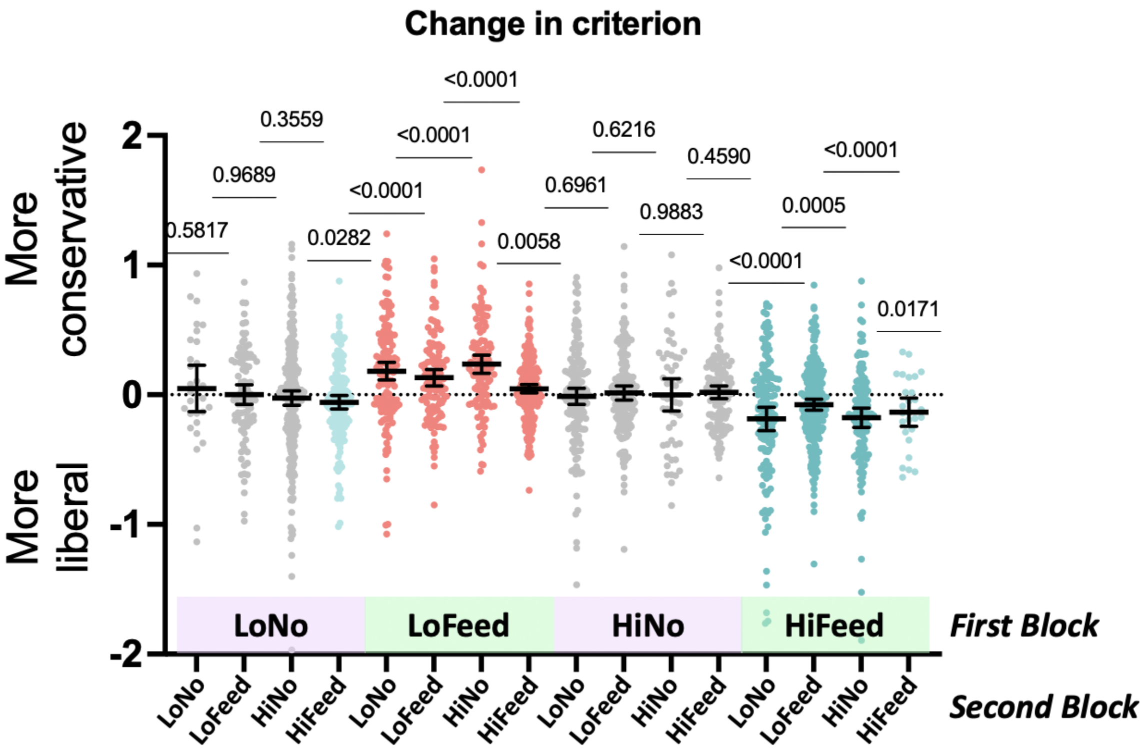


Figure 5: Change in criterion on block 2 as a function of the nature of block 1. P-values show results for simple t-tests, testing against the null hypothesis that there is no change in criterion. Black lines show means +/- 95% confidence intervals.

Looking separately at changes in P(True Positive) and P(False Positive) responses, we see essentially the same pattern of results. For changes in True Positive proportions, pairs where Block 1 is low prevalence with feedback are significant (all $p < 0.005$) except when Block 2 is High Feedback ($p = 0.38$). Pairs where Block 1 is high prevalence with feedback are significant (all $p < 0.005$ except when Block 2 is Low Feedback which $p = 0.02$). The Low Feedback – High Feedback pair is not significant. For changes in False Positive proportions, pairs where Block 1 is low prevalence with feedback are significant (all $p < 0.05$). When Block 1 is high prevalence with feedback, results are significant when Block 2 is Low No Feedback, High No Feedback, and Low Feedback ($p < 0.01$). When Block 2 is High Feedback, there is no significant change ($p = 0.29$). The Low No Feedback – High Feedback pair is significant ($p = 0.001$). No other changes in P(True Positive) and P(False Positive) responses are significant. As noted above, removing the d' > 0.5 filter does not change the pattern of results. Limiting analysis to only the first pair of blocks for each participant preserves the direction of effects but some effects become statistically unreliable because of the loss of power.

Do the effects of one block on the next change with time between blocks?

Figure 5 shows that a block with feedback has an influence on the criterion in the next block. Is that effect transitory or persistent? The time between blocks is quite variable because participants could perform the second block immediately or from one to several days later. Accordingly, for each of the 16 pairs of blocks, we examined change in criterion as a function of time between blocks. The results are shown for four different time ranges in Figure 6.

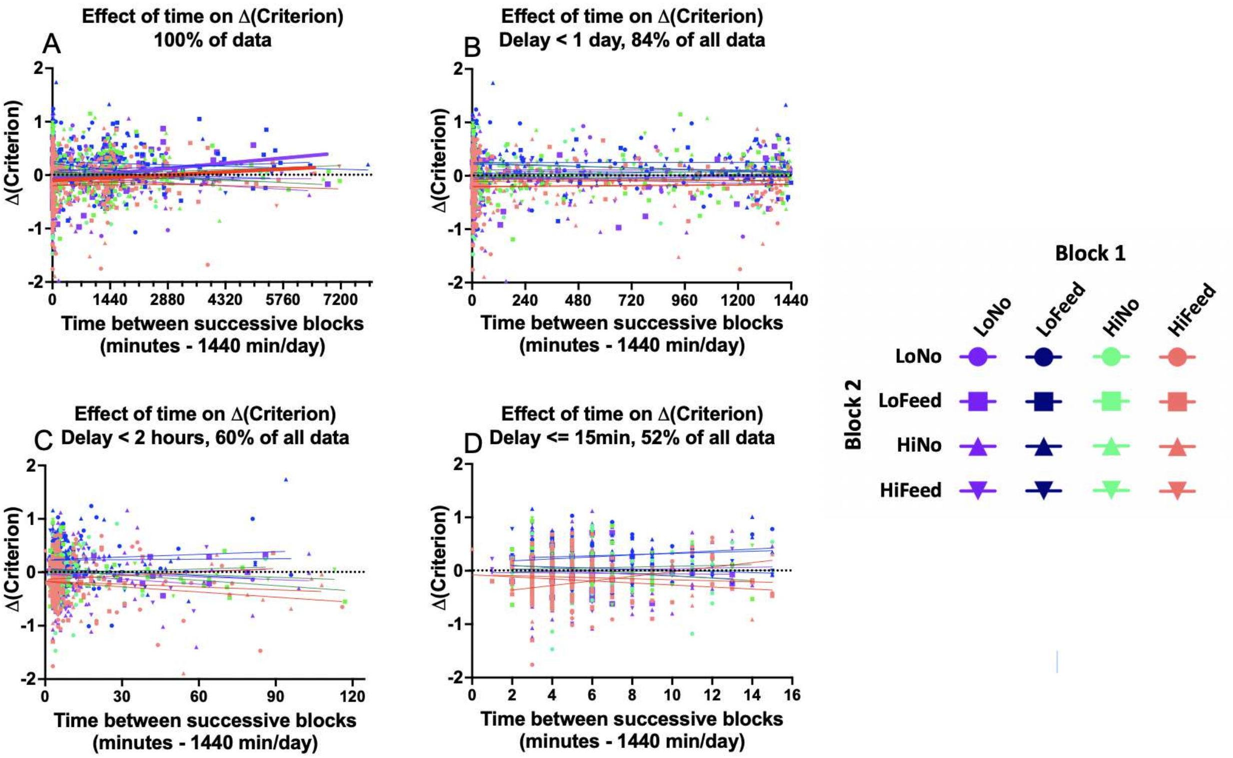


Figure 6: Change in criterion as a function of time between the start of block 1 and start of block 2. 1440 minutes = 1 day. Colored lines are best-fit linear regressions for each of 16 pairs of block types. A) All data, B) Delay < 1 day, C) Delay < 2 hours, D) Delay < 15 min. Repeat pairs (e.g. Low Feedback -> Low Feedback) do not occur for shorter delays (C & D).

Because the distribution of delays is strongly skewed toward shorter delays, we repeated the regression analysis separately for the whole dataset, for all delays < 1 day (84% of all data), < 2hrs (60% of all data), and 15 min or less (still include 52% of all data because participants tended to do one block and then another, immediately). The four panels of Figure 6 zoom in to smaller and smaller time scales but, in fact, it does not matter. The point of the admittedly noisy Figure 6 is that there is no obvious pattern of slopes. This is borne out by linear regressions for each of the 16 pairs of block types. For the entire data set (Figure 6A), fourteen of sixteen correlations are not significant (all 14 r-sq < 0.02, all p > 0.16). The Low No Feedback -> Low Feedback Pair has a r-sq of 0.06 (p = 0.015). The High Feedback -> Low Feedback Pair has a r-sq of 0.02 (p = 0.045). The regression lines for those conditions are shown as thicker lines in Fig 6A. Examining this relationship, the change in criterion is near zero at short times and the change grows more conservative with time. For delays less than 1 day, no correlations are significant (all r-sq < 0.02, all p > 0.35). For delays less than 2 hrs, no correlations are significant (all r-sq < 0.02, all p > 0.20). For delays less than or equal to 15 min, no correlations are significant (all r-sq < 0.025, all p > 0.24). In sum, there is no evidence for the effects, shown in Figure 4, fading with time. The two (out of 64) significant correlations seem likely to be random fluctuations of the data and go in the 'wrong' direction if the hypothesis under test is that block 1 would influence block 2 if block 2 occurred recently but not after a longer delay.

This result is somewhat surprising since, surely, the impact of one block on the next must fade at some point. The results, shown in Figure 6 tell us that the fading is not fast. In this, the effect of one block seems more like education than adaptation. The first block in the pair is teaching participants something and, like learning the capital of Sweden, that knowledge does not simply fade away within minutes or hours. One might object that the results shown in Figure 6 are, essentially, negative. As a different way to show that the effects are persistent, Figure 7 replicates Figure 5 but with only 40% of the data included, the 40% with block 1 – block 2 delays longer than two hours. The effects get somewhat weaker since most of the data has been discarded but it is clear that the pattern remains the same. Exposure to a block of low prevalence with feedback makes participants more conservative on the next block. Exposure to a block of high prevalence with feedback makes participants more conservative.

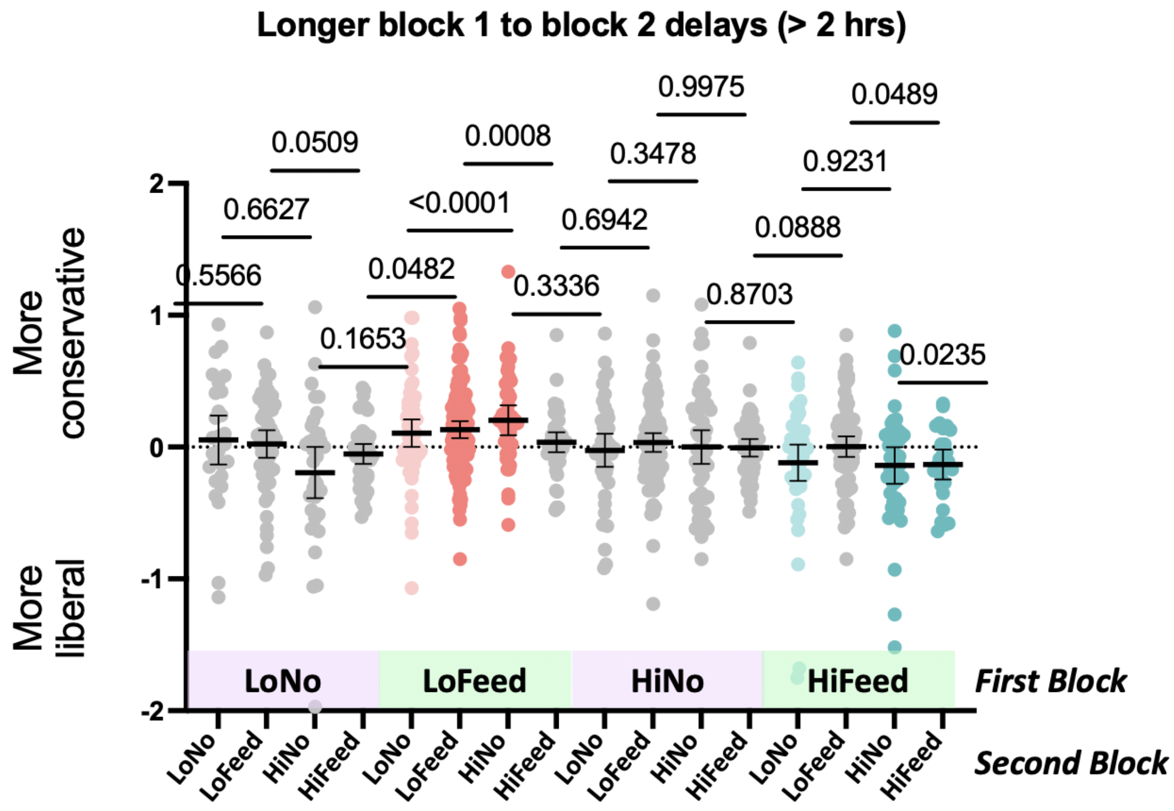


Figure 7: Change in criterion on block 2 as a function of the nature of block 1. Data restricted to Pairs separated by more than 2 hours.

Does the effect of one block last throughout the next block?

Wolfe and VanWert (2010) did visual search experiments in which target prevalence changed smoothly over 1000 trials. Criterion tracked the change in prevalence with a lag that suggested that criterion was based on the last 2-3 dozen trials. Thus, in an 80-trial block, one might expect the effect of block 1 to be present in the first 40 trials and, perhaps, reduced or absent in the second 40 trials. To assess that possibility, the analysis, shown in Figure 5 was repeated for the first half and second half of the trials in block 2, separately. The results are essentially unchanged from those shown in Figure 5. Exposure to Low Prevalence with Feedback in block 1, produces a conservative shift in block 2 criterion in both the first half and the second half of block 2. All t-tests for pairs where Block 1 is Low Feedback are significant except for the Low Feedback → High Feedback pair in the first half, when the effect should be stronger than in the second half. Exposure to High Prevalence with Feedback in block 1, produces a liberal shift in block 2 criterion in both the first half and the second half of block 2. T-tests when block 2 is Low Feedback, Low No Feedback and High No Feedback are significant. T-test when block 2 is High Feedback is not significant in either half (again, recall that the High Feedback -> High Feedback pair has the fewest instances in the dataset).

The lack of any systematic decrease in the magnitude of the effects is shown in Figure 8. The mean change in criterion is plotted for the first and second half of each pair of blocks. Note the grouping of the pairs. The dark blue pairs show the conservative (positive) shift following low prevalence with feedback. The red pairs show the liberal (negative) shift following high prevalence with feedback. If the effects only lasted for the first 2-3 dozen trials, these effects should collapse toward zero in the second half. Clearly, this is not the case.

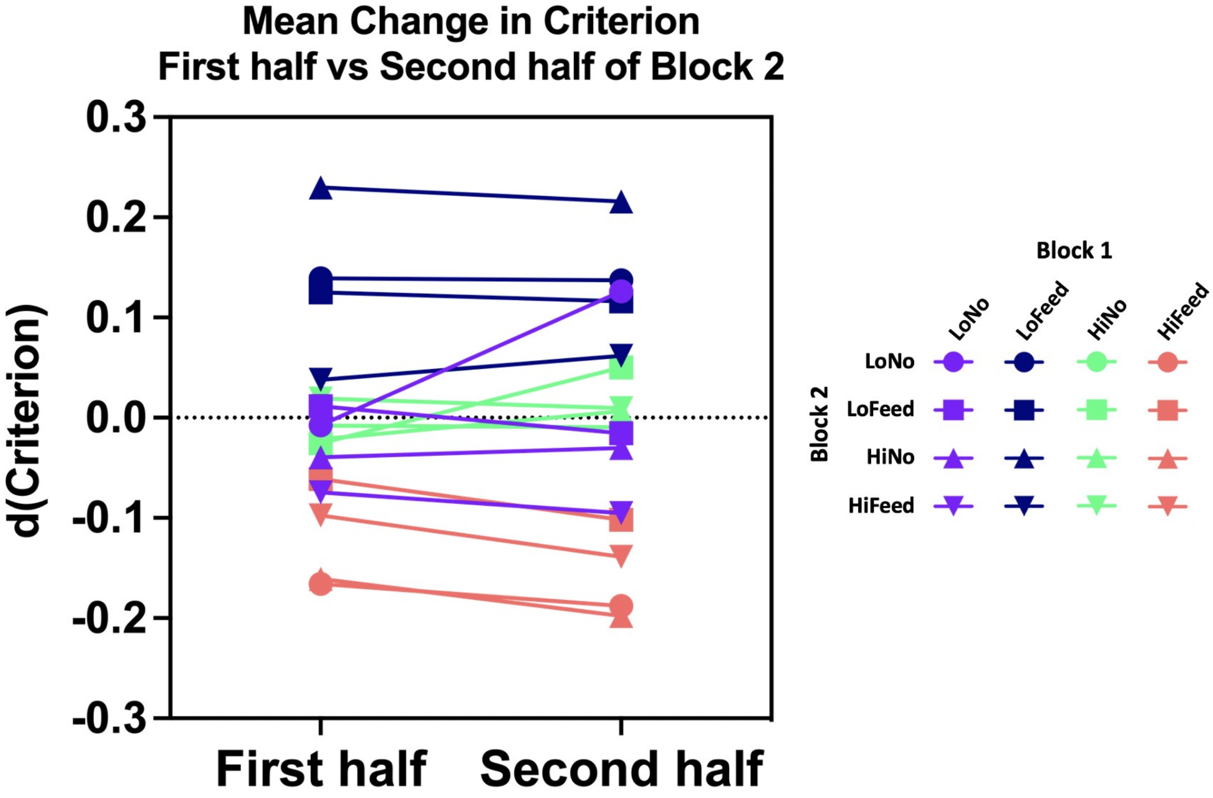


Figure 8: Mean change in criterion for the first and second halves of a block of trials.

Effects of Expertise

As noted in the Methods section, participants gave a rough categorization of their level of expertise. Sadly, this convenience population does not include any substantial number of self-identified expert dermatologists. Only one MD self-identifies as a dermatologist. We can create an expertise continuum from participants reporting no medical training to premedical students to medical students and, finally, to medical doctors. Table 6 shows the numbers of participants and the number of pairs of data contributed by each group.

Expertise category	No. of Participants	No. of pairs
Medical Doctor	54	131
Medical Student	337	952
Pre-Med Student	174	426
No Medical Experience	145	358

Table 6: Number of participants in each expertise group and the number of condition pairs contributed by the groups.

Figure 9 shows d’ and criterion as a function of expertise group, regardless of the type of feedback or prevalence (Recall, from Figure 2, that there is little effect of block type on d’.)

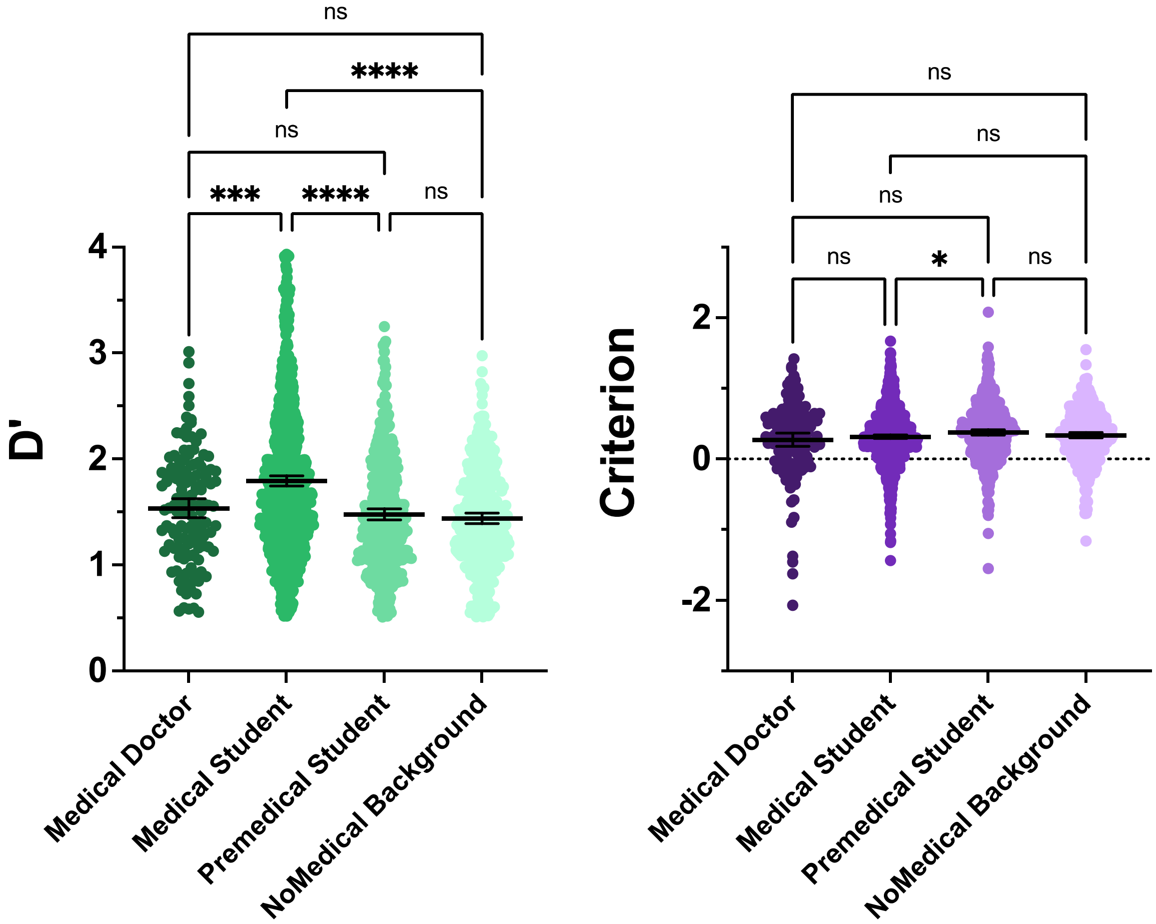


Figure 9: Effects of expertise category on d’ and crit, regardless of prevalence and/or feedback.

There is a highly significant effect of expertise category (F(3, 1863) = 38.58, p < 0.0001, partial eta-squared = 0.86) and, as can be seen by the pairwise comparisons, that effect is entirely due to superior performance of the medical students (all comparisons to medical students, p <= 0.0001). It might seem disappointing that MDs were no better at this task than non-medical participants. However, these non-dermatologist MDs may not have spent time looking at skin lesions for many years and the novices were only being asked to perform a difficult, but straight-forward, two-alternative, forced-choice discrimination. It is the medical students, many of whom may have recently been learning about skin lesions, who do somewhat better on this task. Of course, there could be other reasons for the group differences (e.g. motivated, competitive students versus more casually interested MDs. We cannot know in this case.)

There is a very modest effect of expertise category on criterion (F(3, 1863) = 3.32, p = 0.02, partial eta-squared = 0.01). Pairwise comparisons reveal that premedical students are slightly more conservative than medical students (p = 0.049). This does not appear to be a particularly meaningful finding.

In terms of the impact of expertise on the effect of block 1 on block 2, the data are noisy, with fewer pairs showing significant effects of block 1 on block 2. This seems to be a statistical power issue. The pairs that are significant are the same as those seen in Figure 4, with low prevalence on block 1 making participants more conservative on block 2 and high prevalence making them more liberal. There is no evidence for any systematic effect of expertise. In particular, there is no evidence that the most expert group (medical students) shows a different pattern of results. In that group (fortunately, the largest), Pairs with Low Feedback as block 1 are more conservative and with High Feedback as block 1 are more liberal (all p <= 0.01).

Literature Cited

Codella, N., Rotemberg, V., Tschandl, P., Celebi, M. E., Dusza, S., Gutman, D., et al. (2019). Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC). *arXiv:1902.03368*.

Evans, K. K., Birdwell, R. L., & Wolfe, J. M. (2013). If You Don't Find It Often, You Often Don't Find It: Why Some Cancers Are Missed in Breast Cancer Screening. . *PLoS ONE 8(5): e64366. , 8*(5), e64366. doi: doi:10.1371/journal.pone.0064366

Evans, K. K., Tambouret, R., Wilbur, D. C., Evered, A., & Wolfe, J. M. (2011). Prevalence of Abnormalities Influences Cytologists' Error Rates in Screening for Cervical Cancer" *Archives of Pathology & Laboratory Medicine, 135*(12), 1557-1560. . doi: doi: 10.5858/arpa.2010-0739-OA.

Evered, A. (2017). The prevalence problem in the era of human papillomavirus screening. *Cytopathology, 29*, 97-99.

Growns, B., & Kukucka, J. (2021). The prevalence effect in fingerprint identification: Match and non-match base-rates impact misses and false alarms. [https://doi.org/10.1002/acp.3800]. *Applied Cognitive Psychology, 35*(3), 751–760. doi: https://doi.org/10.1002/acp.3800

Horowitz, T. S. (2017). Prevalence in Visual Search: From the Clinic to the Lab and Back Again. *Japanese Psychological Research, 59*(2), 65-108. doi: 10.1111/jpr.12153

Jackson, S. L., Cook, A. J., Miglioretti, D. L., Carney, P. A., Geller, B. M., Onega, T., et al. (2012). Are Radiologists' Goals for Mammography Accuracy Consistent with Published Recommendations? *Academic Radiology, 19*(3), 289-295. doi: 10.1016/j.acra.2011.10.013

Levari, D. E., Gilbert, D. T., Wilson, T. D., Sievers, B., Amodio, D. M., & Wheatley, T. (2018). Prevalence-induced concept change in human judgment. *Science, 360*(6396), 1465-1467. doi: 10.1126/science.aap8731

Lyu, W., Levari, D. E., Nartker, M., Little, D. S., & Wolfe, J. M. (2021). Feedback moderates the effect of prevalence on perceptual decisions. *Psychon Bull Rev, 28*, 1906–1914. doi: https://doi.org/10.3758/s13423-021-01956-3

Mitroff, S. R., Ericson, J. M., & Sharpe, B. (2017). Predicting airport screening officers' visual search competency with a rapid assessment. *Human Factors, 60*(2), 201-211. doi: https://doi.org/10.1177/0018720817743886

Papesh, M. H., Heisick, L. L., & Warner, K. A. (2018). The persistent low-prevalence effect in unfamiliar face-matching: The roles of feedback and criterion shifting. *Journal of Experimental Psychology: Applied, 24*(3), 416-430. doi: 10.1037/xap0000156

Reed, W. M., Ryan, J. T., McEntee, M. F., Evanoff, M. G., & Brennan, P. C. (2011). The Effect of Abnormality-Prevalence Expectation on Expert Observer Performance and Visual Search. *Radiology, 258*(3), 938-943. doi: radiol.10101090 [pii]
10.1148/radiol.10101090

Trueblood, J. S., Eichbaum, Q., Seegmiller, A. C., Stratton, C., O'Daniels, P., & Holmes, W. R. (2021). Disentangling prevalence induced biases in medical image decision-making. *Cognition, 212*, 104713. doi: https://doi.org/10.1016/j.cognition.2021.104713

Weatherford, D. R., Erickson, W. B., Thomas, J., Walker, M. E., & Schein, B. (2020). You shall not pass: how facial variability and feedback affect the detection of low-prevalence fake IDs. *Cognitive Research: Principles and Implications, 5*(1), 3. doi: 10.1186/s41235-019-0204-1

Wolfe, J. M., Brunelli, D. N., Rubinstein, J., & Horowitz , T. S. (2013). Prevalence effects in newly trained airport checkpoint screeners: trained observers miss rare targets, too. *J of Vision, 13*(3), 33. doi: 10.1167/13.3.33

Wolfe, J. M., Horowitz, T. S., & Kenner, N. M. (2005). Rare targets are often missed in visual search. *Nature, 435*(7041), 439-440. doi: 10.1038/435439a

Wolfe, J. M., Horowitz , T. S., VanWert, 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*(4), 623-638. doi: doi: 10.1037/0096-3445.136.4.623.

Wolfe, J. M., & Van Wert, M. J. (2010). Varying Target Prevalence Reveals Two Dissociable Decision Criteria in Visual Search. *Curr Biol, 20*(2), 121-124. doi: https://doi.org/10.1016/j.cub.2009.11.066