Psychonomic Bulletin & Review
https://doi.org/10.3758/s13423-021-01956-3

BRIEF REPORT

Feedback moderates the effect of prevalence on perceptual decisions

Wanyi Lyu[1] · David E. Levari[3] · Makaela S. Nartker[4] · Daniel S. Little[5] · Jeremy M. Wolfe[1,2]

Accepted: 17 May 2021

Abstract

How does the prevalence of a target influence how it is perceived and categorized? A substantial body of work, mostly in visual search, shows that a higher proportion of targets are missed when prevalence is low. This classic low prevalence effect (LPE) involves a shift to a more conservative decision criterion that makes it less likely that observers will call an ambiguous item a target. In contrast, Levari et al. (*Science, 360*[6396], 1465–1467, 2018) recently reported the opposite effect in a simple categorization task. In their hands, at low prevalence, observers adopted a more liberal criterion, making observers more likely to label ambiguous dots on a blue–purple continuum "blue." They called this "prevalence-induced concept change" (PICC). Here, we report that the presence or absence of feedback is critical. With feedback, observers become more conservative at low prevalence, as in the LPE. Without feedback, they become more liberal, identifying a wider range of stimuli as targets, as in Levari's PICC studies. Stimuli from a shape continuum ranging from rounded ("Bouba") to bumpy ("Kiki") shapes produced similar results. Other variables: response type (2AFC vs. go/no-go), color (blue–purple vs. red–green), and stimuli type (solid color vs. texture) did not influence the criterion shifts. Understanding these effects of prevalence and ways they can be controlled illuminates the context-specific nature of perceptual decisions and may be useful in socially important, low prevalence tasks like cancer screening, airport security, and disease diagnosis in pathology.

Keywords Prevalence effects · Decision · Criterion · Categorical perception

Many important tasks involve decisions about rare items (e.g., detecting threats in luggage or identifying tumors in X-rays). Unfortunately, expert and nonexpert observers are less likely to find or identify targets when they are rare (Biggs et al., 2014; Colquhoun & Baddeley, 1967; Evans et al., 2013; Wolfe et al., 2013; Wolfe et al., 2005; reviewed in Horowitz, 2017). In contrast, Levari et al. (2018) found that people categorizing colored dots, threatening faces, or unethical scientific experiments became more liberal in labeling ambiguous items as targets at low prevalence. What determines whether observers' decision criteria become more liberal or more conservative when target prevalence is low? This paper shows that the presence or absence of feedback is one explanation for this apparently contradictory set of findings.

Studies of vigilance and visual search have shown that observers are less likely to detect targets as prevalence decreases, and that *fewer* ambiguous stimuli are labeled as targets. For instance, observers in Wolfe et al. (2007) searched for guns and knives in a simulated airport X-ray baggage-screening task. Miss errors increased considerably when target prevalence decreased from 50% to 2%. Signal detection analysis showed that the primary effect of low prevalence was a conservative criterion shift that rendered the observer less likely to declare ambiguous stimuli to be targets. There have been many replications of what we will call the classic low prevalence effect (LPE) in visual search tasks (reviewed in Horowitz, 2017). Similar results are found in vigilance tasks where observers must respond to intermittent signals over an extended period of time (Warm, 1993; Warm et al., 2015). Again, low prevalence is associated with higher miss rates (Baddeley & Colquhoun, 1969; Colquhoun & Baddeley, 1964, 1967; Thomson et al., 2016).

✉ Wanyi Lyu
wlu3@bwh.harvard.edu

[1] Brigham & Women's Hospital, Boston, MA, USA
[2] Harvard Medical School, Boston, MA, USA
[3] Harvard Business School, Boston, MA, USA
[4] Johns Hopkins University, Baltimore, MD, USA
[5] Bowdoin College, Brunswick, ME, USA

Published online: 25 June 2021

One way of thinking about the LPE is that it represents a narrowing of the definition of a target. In contrast, Levari et al. (2018) found that low prevalence produced a broadening of the target definition. In one of their experiments, observers judged whether a colored dot, drawn from a blue–purple continuum, was blue or not blue on each trial (see Fig. 1a). The probability of a dot coming from the blue half of the continuum started at 50% and declined gradually to 6% over the course of the experiment. At low prevalence, observers were more likely to call ambiguous dots "blue." Levari et. al dubbed this liberal criterion shift "prevalence-induced concept change" (PICC).

In the PICC experiments, observers are making decisions about a single stimulus at a time, while in most recent LPE studies, the task has been visual search for a target amongst distractors. However, pilot experiments in our lab as well as the older vigilance work show that the LPE can also be produced with single stimuli (Baddeley & Colquhoun, 1969; Colquhoun & Baddeley, 1967). Thus, the difference between LPE and PICC does not seem to be due to the nature of the stimuli. Instead, here we focus on the information available to observers in these tasks. Observers can keep track of two variables in simple, two-alternative forced-choice (2AFC) decision tasks. They know what answer they made and, if feedback is provided, they know if they got the answer right. As target prevalence decreases, an observer can notice the decreasing rate of correct "yes/target present" (hit) responses, and the increasing relative frequency of false positive (false

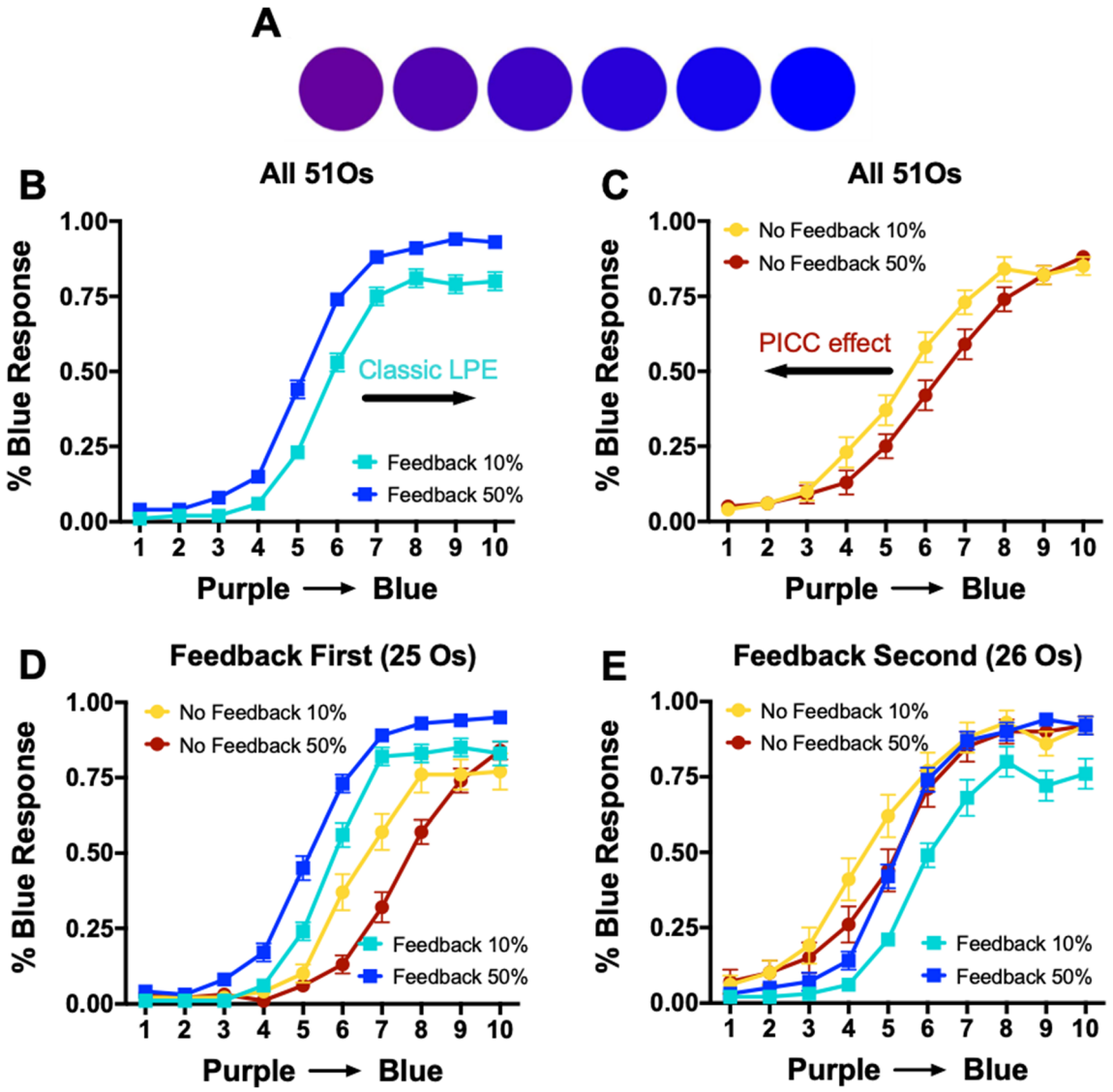

Fig. 1 Stimuli and Results for Experiment 1. a The blue–purple stimulus continuum. **b** The percentage of blue responses as a function of the color category and target prevalence (50%/10%) from the *feedback* condition for all 51 Os. **c** Results from the *no-feedback* condition. **d** Results from both feedback conditions when observers ran *feedback* condition first (25 Os). **e** Results from both feedback conditions when observers ran *feedback* condition second (26 Os). Error bars are ±1 *SEM*

alarm) errors. That is, when they make a mistake at low prevalence, it becomes more likely that the mistake will be a false positive.

If observers aim (implicitly or explicitly) to behave similarly at low and high prevalence, these two types of information will exert opposite pressures on decisions. An observer monitoring *response type* may notice that low prevalence makes them say "no" more often and might respond by increasing "yes" responses. This would be pressure in the PICC direction. Alternatively, especially with clear feedback, an observer monitoring *errors* could notice that low prevalence makes them produce too many false alarms. They might respond by decreasing "yes" responses. This would be pressure in the LPE direction. In other words, which kind of information is monitored—responses or errors—may determine how prevalence influences target detection (see Vickers & Leary, 1983).

While observers can usually maintain a rough sense of their rate of "yes" responses in a detection task, monitoring errors is difficult (if not impossible) unless feedback is given about accuracy. In the present study, when we manipulated the presence or absence of feedback, we found that feedback was crucial in determining whether observers produce a classic LPE or a PICC effect in perceptual decisions on both color and shape continua. Control experiments tested the effect of response type (2AFC vs. go/no-go), color (blue–purple vs. red–green), stimulus type (solid color vs. texture), and presence or absence of exemplars. These factors did not influence criterion shift (see supporting information). Our results suggest that low target prevalence can put powerful, opposing pressures on response criteria for perceptual decisions.

Experiment 1

Experiment 1 investigated whether the presence of feedback changes response behavior at low target prevalence. In Levari et al. (2018), observers judged whether a colored dot, drawn from a blue–purple continuum, was blue or not blue on each trial (see Fig. 1a). No feedback was provided to the observers after their decisions on each trial. We hypothesized that in the absence of feedback, as blue dots became rarer, observers might have noticed themselves making fewer target-present responses, but could not be sure whether the number of targets was actually declining. Observers might respond to such an observation by increasing the number of target-present responses, producing a PICC effect. However, if feedback is given at low prevalence, observers might learn that they are incorrectly labeling more nontargets as targets, and, in an effort to decrease the high false positive rate, they might reduce the number of positive responses, producing a classic LPE. Experiment 1 tests this prediction by manipulating the availability of trial-by-trial feedback.

Methods

Preregistration

Experiment 1 was preregistered on the Open Science Framework (https://osf.io/cgy4p/), where raw data files are publicly available. We stated that observers needed to perform at 70% correct in order to be included in the analysis. "Correct" in this case is defined by an arbitrary division of the purple–blue continuum into blue and not blue. The 70% criterion merely assures that an observer was generating a meaningful function relating judgements to stimulus color and not, for instance, replying randomly.

Observers and power

In a pilot version of the experiment (Exp. S3 in SI), we obtained an LPE with an effect size of 1.2 (Cohen's *d*) with feedback and a PICC with an effect size of 0.6 without feedback. To detect a shift of a neutral point on the psychometric function with an effect size of 0.6, alpha = 0.01, and power = 0.8 requires 31 observers. Given uncertainties with online data quality, we tested 60 observers using the Amazon Mechanical Turk (MTurk) online platform. Individuals located in the United States were invited to participate as long as their MTurk approval rate was above 95%. We discuss the exclusions of observers below. Observers were paid $6/hour. Procedures were approved by the Institutional Review Board at Brigham and Women's Hospital (IRB #2007P000646). All observers were naïve to the purpose of the experiment.

Stimuli and procedure

The stimuli and procedure were chosen to closely resemble those of Levari et al. (2018). The basic requirement is a unidimensional stimulus continuum that runs from stimuli that are clearly not targets, through an ambiguous region, to stimuli that are clearly targets. On each trial in this experiment, a dot stimulus was presented at the center of the screen. The dot had a radius of 200 pixels, equivalent to 5 degrees of visual angle at a viewing distance of approximately 60 cm. Dots appeared on a solid white background. The colors of the dots were drawn from a blue–purple continuum with 100 discrete RGB values (most purple: RGB 100-0-155, CIExy 0.232, 0.115; most blue: RGB 1-0-254, CIExy 0.143, 0.051). We divided the color spectrum into two halves that we referred to as the "blue distribution" (RGB 50-0-205 through 1-0-254) and the "non-blue (purple) distribution" (RGB 100-0-155 through 51-0-204; The CIExy coordinates for the midpoint are 0.159, 0.063). The stimuli are not equiluminant (blue: 14.3 cd/$m^{2,}$ purple: 13.5 cd/m^2). Since this was run online, the color values must be considered an approximation.

At the beginning of the experiment, observers were told that they would see a series of colored dots and were instructed to press one key if they considered a dot to be blue, and another key if it was not blue. The choice of a blue/not-blue decision follows the methods Levari et al. (2018). In other studies, we have used other responses (e.g., more red/more green). The results are essentially the same (see Supplementary Information Exp. S3). The dots were presented on the screen one at a time for 500 ms. Observers were told that some series of dots could include many blue dots while others might have only a few. They were asked to answer as quickly and accurately as possible. Each observer completed two conditions: *feedback* and *no feedback*. One group of observers ($n = 24$) received the *feedback* condition first while the other group received the *no-feedback* condition first. Each condition consisted of 600 trials. The prevalence of targets declined in an abrupt but unmarked step from 50% directly to 10%. This differs from the gradual decline from 50% to 6% in Levari et al. (2018). However, in Levari et al. Study 4, they replicated their PICC effect with an abrupt change from 50% to 6%. We set the low prevalence rate to 10% in order to increase the number of low prevalence trials over the numbers obtained with 6% prevalence in the previous experiments. Thus, in each feedback condition, we drew 50% of the dots from the blue distribution for trials 1–200, and only 10% from the blue distribution over trials 201–600.

Observer exclusions

The data consist of psychometric functions relating the proportion of blue responses to the color of the dot (see Fig. 1). We removed trials with reaction time shorter than 200 ms or longer than 3,000 ms. Three observers with less than 20% valid trials were excluded. We excluded five observers who had completely flat psychometric functions in one or more conditions. They had either all blue or all not-blue responses. There is an intermediate set of observers who have one or more very shallow functions despite that their overall accuracy is above the exclusion criterion. The difficulty is that we have no a priori way to know whether these functions reflect careless and inconsistent performance of the task or—for example, dramatic low prevalence effects. Accordingly, we note this group but do not exclude it. We define this group as consisting of any observers with two or more functions that never reach over 80% "blue" responses (meaning, they called the bluest of blue items "not blue" at least 20% of the time). We also included observers who had a single function that never rose above 60% blue responses. There are 12 observers in this intermediate group. There are 39 observers without problematic psychometric functions. Exclusions reduce the noise in the data but do not change the pattern of results.

Results

Figures 1b–c plot the proportion of blue responses as a function of binned stimulus color for 10% and 50% blocks, separately for *feedback* and *no-feedback* conditions. Data are shown combined for all 51 observers in the unproblematic (39 observers) and intermediate (12 observers) groups. Colors were binned into 10 color categories (10 most blue, 1 least blue). As is evident in Fig. 1b, with feedback, there is a shift to a lower percentage of blue responses for all colors in the 10% prevalence blocks. That appears as a shift of the 10% curve to the right of the 50% curve, corresponding to the traditional LPE. Importantly, without feedback (see Fig. 1c), the low prevalence function shifts to the left, producing a PICC effect (a higher percentage of blue responses). If analyzed separately, the results for the 39 unproblematic observers look very similar to Fig. 1, while the 12 "intermediate" observers produce shallower (noisier) functions where the leftward shift in Fig. 1c is less obvious.

A logistic regression with prevalence and feedback as factors in a generalized mixed model run using *jamovi* (Jamovi Project, 2020) shows that both prevalence and feedback factors are significant (Feedback: $\chi^2 = 793$, $df = 1$, $p < .001$; Prevalence: $\chi^2 = 1127$, $df = 1$, $p < .001$). The prevalence effect goes in one direction when there is feedback and the other direction when there is not. The interaction is significant ($\chi^2 = 1268$, $df = 1$, $p < .001$). Even if the analysis is restricted to the 12 intermediate group observers, the main effects and the interaction remain significant (all ps $< .001$).

The signal detection measures of d' and criterion (c) can be computed by dividing the continuum into two halves and defining as false alarms any blue responses to Categories 1–5. Blue responses to Categories 6–10 are defined as true positive (hit) responses. The detectability of the signal is measured by d', and c measures bias or criterion. Statistical details are shown in the Supporting Information (Fig. S6). In brief, for each feedback condition, d' remains largely unchanged while c becomes more conservative with feedback and more liberal, without feedback.

Did the order of feedback conditions matter? Some observers in this experiment experienced the *feedback* condition followed by the *no-feedback* condition, while others had the reverse. Figure 1d–e show that the important pattern of results is unaffected by the order: PICC effect without feedback, LPE with feedback. A closer look at the two figures indicates that when *feedback* condition comes second (Fig. 1e), the two 50% prevalence functions roughly align with each other, suggesting that observers hold the same response criterion at the start of the two feedback conditions. Interestingly, when *feedback* condition comes first (Fig. 1d), the two *no-feedback* functions (red/yellow) shift to the right of the two *feedback* functions (blue/cyan). In addition, the % Blue Response for the two *no-feedback* functions drops for the most clearly blue

colors (8–10), as happens in the *feedback* 10% function. Thus, when *no-feedback* blocks follow the *feedback* blocks, observers seem to retain the conservative decision criteria that they adopted in the immediately preceding 10% prevalence *feedback* trials. The main effect of feedback order is not significant (generalized mixed model: $\chi^2 = 1.35$, $df = 1$, $p = .25$). The effect of feedback order appears in interactions with feedback ($\chi^2 = 1294$, $df = 1$, $p < .001$) and prevalence ($\chi^2 = 5.40$, $df = 1$, $p < .05$). Adding feedback order somewhat improved the generalized mixed model (AIC: 33705 vs. 35955 for the previous model, where smaller is a better model).

The average psychometric function in the *no-feedback* condition is shallower than the average function with feedback. This arises because, without feedback, observers are free to define "blue" as they see fit. Consequently, individual functions vary in their horizontal position. When averaged together, they produce a fairly shallow function. With feedback, the experimenters define "'blue" and the observers all produce similar functions. The fact that d' is not changed by feedback shows that feedback is not altering discriminability and that the slope of the average function does not reflect individual d'.

Before discussing the implications of this result, we present a replication, using a different type of stimulus.

Experiment 2

Levari et al. (2018) showed similar PICC effects using other types of stimuli such as faces and experiment proposals. Likewise, the classic LPE has been demonstrated with tasks such as cancer and threat detection (Evans et al., 2013; Wolfe et al., 2005). Clearly, change in the decision criterion is not limited to judgements pertaining to color alone. Experiment 2 aimed to determine whether the effects of feedback in Experiment 1 generalize to other stimuli. We used a continuum of "Bouba–Kiki" shape stimuli, as shown in Fig. 2a. The Bouba–Kiki effect was first described by Wolfgang Köhler in 1929. He reported that observers were more likely to associate the word "baluba" with rounder shapes, and "takete" with more angular/pointier shapes (Köhler, 1929). We created our own shapes for this continuum, using the terms "Bouba" and "Kiki" (Ramachandran & Hubbard, 2001).

Methods

Preregistration

Experiment 2 was preregistered on the Open Science Framework (https://osf.io/ets7x/), where raw data files are also publicly available.

Observers

We tested 48 observers in Experiment 2, based on the power calculation in Experiment 1. The observers were recruited on MTurk and tested on the CloudResearch online platform. They were paid $8/hour. The same observer inclusion criteria as in Experiment 1 were applied. All observers were naïve to the purpose of the experiment.

Stimuli and procedure

The procedure was identical to Experiment 1, except for the following changes:

First, the blue–purple dots were replaced by a set of Bouba–Kiki shape stimuli. These are shape stimuli that vary along a shape continuum from Type 1 (Kiki, very bumpy) to Type 10 (Bouba, very rounded); see Fig. 2a. The shapes were created by summing radial frequencies. The number of radial frequency components rises across the continuum. Thus, Type 3 would contain frequencies 1–3, while Type 9 would contain 1–9. The maximum amplitude of each frequency component is 1/freq for each instance. The actual amplitude of each component is a random value between zero and that max. We divided the continuum into two halves. Types 1–5 are deemed to be "Kikis," while Types 6–10 are "Boubas."

Prior to the experiment, observers were given written descriptions of the definitions of Bouba and Kiki as well as examples of Bouba Categories 8 and 10 and Kiki Categories 1 and 3. Observers completed 10 practice and 600 test trials in each condition. On each trial, observers were presented with a shape for 500 ms and asked to judge whether the shape was a Bouba or not. They were instructed to press one key if they decide that the shape was a Bouba and another key if the shape was not a Bouba. We tested *feedback* and *no-feedback* conditions. As in Experiment 1, to examine the effect of feedback order, half of the observers received the *feedback* condition first, while the other half received *no-feedback* condition first. The prevalence of Bouba decreased in a single step from 50% to 10% without informing the observer. Prevalence was 50% for the first 200 trials and 10% for the remaining 400 trials.

Results

Figure 2 shows proportions of Bouba responses as a function of the shape category for 10% and 50% prevalence in *feedback* and *no-feedback* conditions. With trial-by-trial feedback, there is a strong LPE when the prevalence of "Bouba" decreases, regardless of whether observers perform the *feedback* condition first (Fig. 2b) or second (Fig. 2c). The order of feedback conditions has an effect on the *no-feedback* conditions. Without feedback, when the *feedback* condition comes second (Fig. 2c), there is a clear PICC effect as prevalence decreases, replicating Experiment 1. When the *feedback*

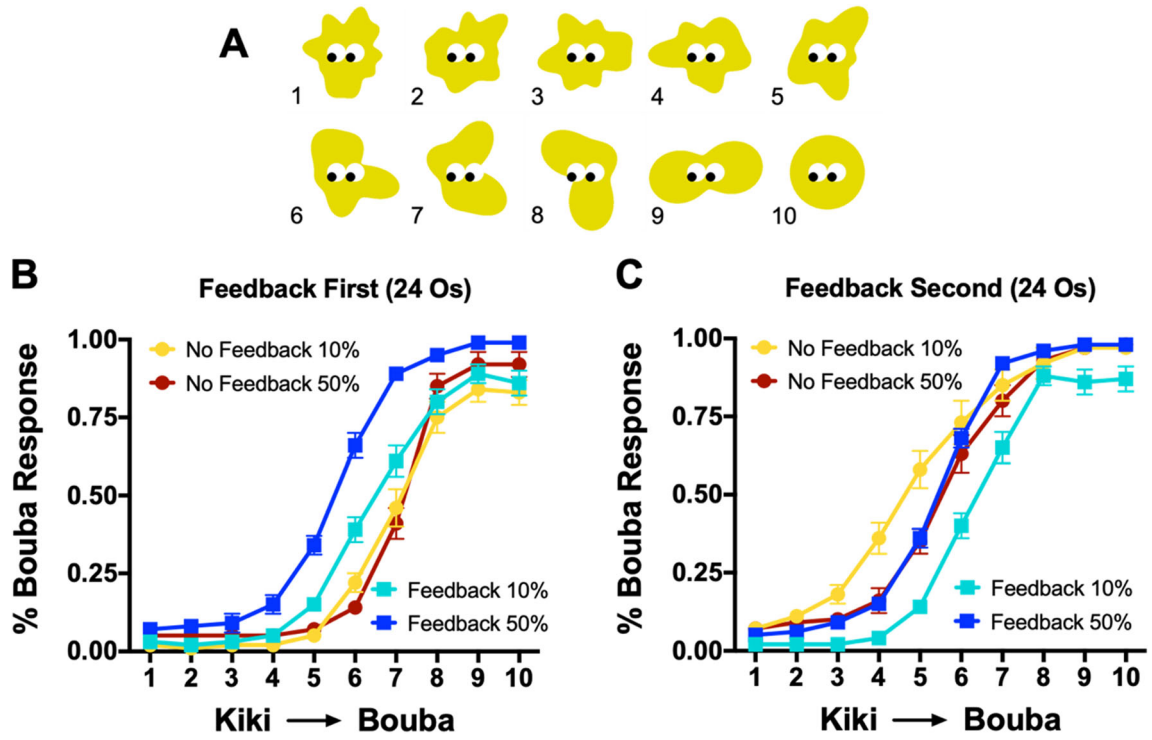

Fig. 2 Stimuli and Results for Experiment 2. a Ten examples of the Bouba–Kiki stimuli with the corresponding shape category values (1–5: Kiki, 6–10: Bouba). **b–c** The percentage of Bouba responses as a function of the shape category and target prevalence (50%/10%), separately for *feedback* and *no-feedback* conditions. **b** Results when observers ran the *feedback* condition first. **c** Results when the observers ran *no-feedback* condition first. Error bars are ±1 *SEM*

condition comes first (Fig. 2b), the *no-feedback* functions resemble the function for the immediately preceding 10% prevalence condition with *feedback,* and there is no PICC effect.

The effects of prevalence and feedback are confirmed with a logistic regression with prevalence, presence of feedback, and feedback order as factors in a generalized mixed model (Feedback: χ^2 = 1870, *df* = 1, $p < .001$; Prevalence: χ^2 = 573.7, *df* = 1, $p < .001$). The prevalence effect differs with and without feedback, producing a significant interaction (χ^2 = 883, *df* = 1, $p < .001$). Again, the main effect of feedback order was not significant (χ^2 = 1.1, df = 1, $p = .30$). It is the interactions of feedback order with feedback (χ^2 = 1413, *df* = 1, $p < .001$) and with prevalence (χ^2 = 4.15, *df* = 1, $p < .05$) that are significant. As in Experiment 1, with feedback, psychometric functions shift to the right (LPE). Without feedback, they either shift left (PICC) or not at all.

As in Experiment 1, we compute c and *d′* for high and low prevalence in the *feedback* and *no-feedback* conditions. The full SDT analysis result is in the Supporting Information (Fig. S7). For the feedback-first data (Fig. 2b), a two-way analysis of variance (ANOVA), with prevalence and feedback as factors, showed significant effects of prevalence and feedback on c and a cross-over interaction (all *p*s < .01). There is no effect of prevalence or feedback on *d′*, nor a cross-over interaction (all *p*s > .07). For the feedback-second data (Fig. 2c), the two-way ANOVA, with prevalence and feedback as factors, shows no significant effect of prevalence on c because the effects are in opposite directions ($p = .10$). There is a substantial main effect of feedback and a substantial cross-over interaction (all *p*s < .001). In *d′*, there is no effect of prevalence, feedback, or a cross-over interaction (all *p*s > .05). A paired *t*-test does show differences in *d′* as a function of prevalence with feedback—feedback first (50% prevalence: 2.59, 10% prevalence: 2.33; paired *t*-test), $t(23) = 3.93$, $p < .001$; and without feedback—feedback second (50% prevalence: 2.51, 10% prevalence: 2.26; paired *t*-test), $t(23) = 3.10$, $p < .01$, reflecting some loss of precision without feedback.

Discussion

Suppose you were manning a medical advice phone line, deciding whether callers reporting their symptoms need to come for a medical test. Sometimes the answer is a clear "yes" or "no." Other times, the data are ambiguous. How do you decide? These experiments and, indeed, the broader literature on prevalence effects, indicate that your decision will be shaped by the current prevalence of disease in your patient population, by any feedback you are getting about your decisions and, perhaps, by your recent training. Effects of low target prevalence have been widely studied in the field of vigilance (Baddeley & Colquhoun, 1969; Colquhoun & Baddeley, 1964, 1967; Thomson et al., 2016), visual search (Horowitz,

2017; Rich et al., 2008), decision-making (Levari et al., 2018; Weatherford et al., 2020), and medical image perception (Evans et al., 2011; Wolfe et al., 2013). In most of that work, lower target prevalence has been accompanied by a more conservative response criterion and a greater tendency to miss targets.

Levari et al.'s (2018) PICC effect goes in the opposite direction, with observers becoming more liberal in their decision criteria. The present studies show that the presence or absence of trial-by-trial feedback is one factor that accounts for this apparent contradiction. With feedback, observers produced LPEs, reducing the rate of target-present responses. Without feedback, observers typically produced PICC effects, increasing the rate of target-present responses. Training with feedback first seems to interfere with the PICC effect in a subsequent, *no-feedback* condition. Interestingly, recent work has shown that fingerprint examiners produce clear LPE with feedback in a fingerprint matching task. They showed no PICC or LPE effect in the absence of feedback (Growns & Kukucka, 2021). Thus, while it seems clear that feedback can strongly modulate the effects of prevalence, explaining the precise effects in different stimulus continua is a topic for future research.

Some other variables are not critical. We evaluated the effect of response type (2AFC vs. go/no-go), stimuli type (solid color vs. texture), color (blue–purple vs. red–green), and the presence and absence of exemplars (reported in Supplementary Information). None of these factors reversed the prevalence effect in the way that feedback did.

How might explicit feedback alter the prevalence effect?

It is important to note that there is some form of feedback in both *feedback* and *no-feedback* conditions. As noted in the introduction, even in the *no-feedback* condition, there can be self-generated feedback about how often each response key was pressed. Changes in prevalence change both explicit, correct/incorrect feedback, and implicit feedback about the ratio of blue to nonblue or Kiki to Bouba responses. When the prevalence of targets and nontargets is equal, observers press the "target" key with some baseline probability. When the prevalence of targets declines, observers will be pushing the "nontarget" button more often. Without explicit feedback, observers might think that their response criterion has shifted to become too conservative. An effort to counteract this illusory error would lead observers to categorize more of the ambiguous stimuli as targets, pushing the criterion to a more liberal position.

With explicit feedback, the error feedback message is more salient and less equivocal than an estimate of how often the "target" key is pressed. When the target prevalence is reduced, false positive errors rise, simply because target-absent trials are far more common. Observers, noting this, might decide (probably without conscious thought) that they need to become more conservative to avoid errors, producing a classic LPE. Both of these shifts can be seen as forms of a base rate problem (Bar-Hillel, 1980). As the prevalence drops, with feedback, observers learn the low base rate of blue/Bouba and respond less often to these targets. In the absence of explicit feedback, observers misperceive the base rate. When observers ran the *no-feedback* condition first, not realizing that it has dropped to 10% when prevalence declines, they attributed the missing target responses to an error that they try to correct. Even when informed that targets might be rare, observers may still improperly account for the changes in the base rate (Bar-Hillel, 1980). When observers ran the no-feedback condition immediately after the low prevalence feedback trials, they may be retaining the base rate that they have acquired based on previous explicit feedback and use it as the baseline for their decisions.

Multiple prevalence effects?

The above account of feedback effects is framed in terms of liberal and conservative shifts of criterion. However, the pattern of results suggests that something more than that is going on. LPE seems to have two manifestations in our data: a shift in the decision criterion for the ambiguous stimuli, and an elevated miss rate even for totally unambiguous stimuli (Category 8–10) in both experiments. At low prevalence, in *feedback* conditions or *no-feedback*—feedback-first condition, the proportion of target-present responses never reached 100%, meaning that some of the most unambiguous target stimuli were not categorized as targets. It could be that observers are making motor errors at low prevalence, simply pushing the nontarget key too quickly (Fleck & Mitroff, 2007), but then we would expect a similar effect in *no-feedback*—feedback-second data (Figs. 1e and 2c). We do not, and motor errors have not proven to be a successful general account for prevalence errors (Hout et al., 2015).

Some of the errors, especially with very clear target stimuli, might be understood as reflections of a speed–accuracy tradeoff. Wolfe and Van Wert (2010) proposed a two-factor account of the LPE, involving a change in a quitting threshold in addition to a criterion shift. Here, too, in both experiments, reaction times are faster at low prevalence, Exp. 1: *feedback*: 31-ms faster, $t(59) = 2.64$, $p < .01$; *no-feedback*: 27-ms faster, $t(59) = 2.12$, $p = .04$; Exp. 2: *feedback*: 57-ms faster, $t(47) = 4.13$, $p < .001$; *no-feedback*: 19-ms faster, $t(47) = 1.34$, $p = .19$. This second prevalence effect may account for the modest decreases in d' seen in some of the low prevalence conditions (see Supporting Information).

Application and next steps

As noted, expert radiologists and TSA workers have been shown to be susceptible to the low prevalence effect, becoming more likely to miss targets when they are rare (Evans et al., 2011;

Wolfe et al., 2013). So, if a clinician moves between a high prevalence setting (e.g., the emergency room) to a low prevalence setting (e.g., a workplace), how might one control unwanted criterion shifts? To reduce a PICC shift in the liberal direction, one obvious answer would be to provide feedback, especially for tasks where the target prevalence changes over time, to minimize the difference between perceived and actual target prevalence. Complete, instant, trial-by-trial feedback is not possible in many real-world situations. Obviously, if one had the information for perfect feedback, it would not be necessary to do the task. It would be interesting to repeat the present experiments under conditions of delayed and/or incomplete feedback.

The idea of incomplete feedback is central to one class of proposed intervention for the case of medical image perception. It might be possible to add a limited number of control cases into an expert's workflow. Since gold-standard truth would be known, feedback could be given right away on these cases. These inserted cases could also serve as a quality assessment/quality improvement (QA/QI) initiative, allowing individuals (and/or their supervisors) to monitor performance. Indeed, Beanland et al. (2014) demonstrated that increasing the prevalence of search targets made them more salient and easier to detect. However, the addition of the positive cases increases the caseload. Moreover, this would be partial feedback, and again, the effects of intermittent feedback are not known. A variation on inserting positive cases into the workflow might be to present observers with a block of high-prevalence trials with feedback before the low prevalence search (Wolfe et al., 2013). This might serve to hold criterion at a roughly neutral position if the effects of training with feedback persisted into a low prevalence work environment, thus potentially counteracting the classic LPE and PICC. Schwark et al. (2012) suggest providing false feedback as a solution. In their study, they added explicit false feedback to increase the observers' perceived number of misses in a simple visual search task. The manipulation rendered the observers more successful in detecting target but at a cost of a higher number of false alarms (see also Littlefair et al., 2017; Reed et al., 2014). Schwark et al. acknowledged that there could be an ethical issue of providing false information to observers in this method and it is not clear that misinformation is a practical solution in the long term. Information that is known to be unreliable is unlikely to maintain its effectiveness.

The clear role of feedback in moderating prevalence effects offers a tool for manipulating decision criteria. Making use of these tools in real-world settings will require further studies of different forms of feedback in real-world tasks.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13423-021-01956-3.

Acknowledgements We thank Chia-Chien Wu for assistance with setting up the online experiment, Jennifer S. Trueblood for advice on performing data analysis. This research was supported by NIH-NEI EY017001.

References

Baddeley, A. D., & Colquhoun, W. P. (1969). Signal probability and vigilance: A reappraisal of the 'signal-rate'effect. *British Journal of Psychology*, *60*(2), 169–178.

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, *44(3)*, 211–233. https://doi.org/10.1016/0001-6918(80)90046-3

Beanland, V., Lenné, M. G., & Underwood, G. (2014). Safety in numbers: Target prevalence affects the detection of vehicles during simulated driving. *Attention, Perception, & Psychophysics*, *76*(3), 805–813. https://doi.org/10.3758/s13414-013-0603-1

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

Colquhoun, W. P., & Baddeley, A. D. (1964). Role of pretest expectancy in vigilance decrement. *Journal of Experimental Psychology*, *68*(2), 156.

Colquhoun, W. P., & Baddeley, A. D. (1967). Influence of signal probability during pretraining on vigilance decrement. *Journal of Experimental Psychology*, *73*(1), 153.

Evans, K. K., Tambouret, R. H., Evered, A., 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(12), 1557–1560. https://doi.org/10.5858/arpa.2010-0739-OA

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), Article e64366. https://doi.org/10.1371/journal.pone.0064366

Fleck, M. S., & Mitroff, S. R. (2007). Rare targets are rarely missed in correctable search. *Psychological Science*, *18*(11), 943–947.

Growns, B., & Kukucka, J. (2021). The prevalence effect in fingerprint identification: Match and non-match base-rates impact misses and false alarms. *Applied Cognitive Psychology, 35*(3), 751–760. 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. https://doi.org/10.1111/jpr.12153

Hout, M. C., Walenchok, S. C., Goldinger, S. D., & Wolfe, J. M. (2015). Failures of perception in the low-prevalence effect: Evidence from active and passive visual search. *Journal of Experimental Psychology. Human Perception and Performance*, *41*(4), 977–994. https://doi.org/10.1037/xhp0000053

Jamovi Project. (2020). Jamovi (Version 1.2) [Computer Software]. Retrieved from https://www.jamovi.org

Köhler, W. (1929). *Gestalt psychology*. Liveright.

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. https://doi.org/10.1126/science.aap8731

Littlefair, S., Brennan, P., Reed, W., & Mello-Thoms, C. (2017). Does expectation of abnormality affect the search pattern of radiologists when looking for pulmonary nodules? *Journal of Digital Imaging*, *30*(1), 55–62.

Ramachandran, V. S., & Hubbard, E. M. (2001). Synaesthesia—A window into perception, thought and language. *Journal of Consciousness Studies*, *8*(12), 3–34.

Reed, W. M., Chow, S. L. C., Chew, L. E., & Brennan, P. C. (2014). Assessing the impact of prevalence expectations on radiologists' behavior. *Academic Radiology*, *9*(21), 1220–1221.

Rich, A. N., Kunar, M. A., Van Wert, M. J., Hidalgo-Sotelo, B., Horowitz, T. S., & Wolfe, J. M. (2008). Why do we miss rare targets? Exploring the boundaries of the low prevalence effect. *Journal of Vision*, *8*(15), 15–15.

Schwark, J., Sandry, J., MacDonald, J., & Dolgov, I. (2012). False feedback increases detection of low-prevalence targets in visual search. *Attention, Perception, & Psychophysics*, *74*(8), 1583–1589. https://doi.org/10.3758/s13414-012-0354-4

Thomson, D. R., Besner, D., & Smilek, D. (2016). A critical examination of the evidence for sensitivity loss in modern vigilance tasks. *Psychological Review*, *123*(1), 70.

Vickers, D., & Leary, J. N. (1983). Criterion control in signal detection. *Human factors, 25*(3), 283–296. https://doi.org/10.1177/001872088302500305

Warm, J. S. (1993). Vigilance and target detection. In B. M. Huey & C. D. Wickens, *Workload transition: Implications for individual and team performance* (pp. 139–170). National Academy Press.

Warm, J. S., Finomore, V. S., Vidulich, M. A., & Funke, M. E. (2015). Vigilance: A perceptual challenge. In M. W. Scerbo (Ed.), *The Cambridge handbook of applied perception research* (pp. 241–283). Cambridge University Press. https://doi.org/10.1017/CBO9780511973017.018

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. https://doi.org/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. *Journal of Vision*, *13*(3), 33–33. https://doi.org/10.1167/13.3.33

Wolfe, J. M., Horowitz, T. S., & Kenner, N. M. (2005). Rare items often missed in visual searches. *Nature*, *435*(7041), 439–440. Agricultural & Environmental Science Collection. https://doi.org/10.1038/435439a

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*(4), 623–638. https://doi.org/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. *Current Biology*, *20*(2), 121–124. https://doi.org/10.1016/j.cub.2009.11.066