Cognition 212 (2021) 104713

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Disentangling prevalence induced biases in medical image decision-making

Jennifer S. Trueblood [a,*,1], Quentin Eichbaum [b], Adam C. Seegmiller [b], Charles Stratton [b], Payton O'Daniels [a], William R. Holmes [c,**,1]

[a] Department of Psychology, Vanderbilt University, USA
[b] Vanderbilt Pathology Education Research Group (VPERG), Department of Pathology, Microbiology and Immunology, Vanderbilt University Medical Center (VUMC), USA
[c] Department of Physics and Astronomy, Vanderbilt University, USA

ARTICLE INFO

Keywords:
Medical image perception
Signal detection theory
Diffusion model
Neural networks
Expertise

ABSTRACT

Many important real-world decision tasks involve the detection of rarely occurring targets (e.g., weapons in luggage, potentially cancerous abnormalities in radiographs). Over the past decade, it has been repeatedly demonstrated that extreme prevalence (both high and low) leads to an increase in errors. While this "prevalence effect" is well established, the cognitive and/or perceptual mechanisms responsible for it are not. One reason for this is that the most common tool for analyzing prevalence effects, Signal Detection Theory, cannot distinguish between different biases that might be present. Through an application to pathology image-based decision-making, we illustrate that an evidence accumulation modeling framework can be used to disentangle different types of biases. Importantly, our results show that prevalence influences both response expectancy and stimulus evaluation biases, with novices (students, $N = 96$) showing a more pronounced response expectancy bias and experts (medical laboratory professionals, $N = 19$) showing a more pronounced stimulus evaluation bias.

1. Introduction

In many real-world decisions, individuals are tasked with identifying rarely occurring targets; TSA baggage screeners examine X-rays to find weapons and pathologists and radiologists inspect medical images (e.g. tissues samples or X-rays) for abnormalities. In both cases, targets (i.e., weapons and abnormalities) rarely occur. Over the past decade, there has been significant interest in understanding decision-making and visual search in low prevalence tasks. A common finding across studies involving extreme prevalence rates (both low and high) is that people (including experts) make more errors when targets are extremely rare or common as compared to equal prevalence cases (Horowitz, 2017; Wolfe et al., 2007; Wolfe, Horowitz, & Kenner, 2005). In particular, low prevalence leads to an increase in misses and high prevalence leads to an increase in false alarms (Wolfe & Van Wert, 2010).

An important question is why does this "prevalence effect" occur? That is, what perceptual or cognitive mechanisms lead to reduced performance at extreme prevalence rates? The standard way of analyzing data from prevalence tasks is using Signal Detection Theory (SDT, Green & Swets, 1966). This approach allows researchers to distinguish between two key parameters in the model: discriminability and criterion. Discriminability measures an individual's ability to distinguish between target present and target absent trials and is generally conceptualized as relating to perceptual ability. The criterion is typically assumed to relate to decision-based processes, reflecting biases in responding. Most studies examining the "prevalence effect" have found that prevalence influences the criterion and not discriminability (Gur et al., 2003; Horowitz, 2017; Wolfe et al., 2005; Wolfe & Van Wert, 2010). This suggests that extreme prevalence rates influence decision biases. However, as demonstrated below, different types of biases can lead to similar criterion shifts in SDT, implying that the model cannot distinguish among various biases that might be present in a given task. Relatedly, in recent research, Witt, Taylor, Sugovic, and Wixted (2015) demonstrated that perceptual effects, in addition to response biases, can influence the criterion without affecting discriminability. Given how commonly SDT is used to study decision tasks, this is a critical weakness and limits understanding of the

* Correspondence to: J. S. Trueblood, Department of Psychology, Vanderbilt University, Nashville, TN 37235, USA.
** Correspondence to: W. R. Holmes, Department of Physics and Astronomy, Vanderbilt University, Nashville, TN 37235, USA.
E-mail addresses: jennifer.s.trueblood@vanderbilt.edu (J.S. Trueblood), quentin.eichbaum@vanderbilt.edu (Q. Eichbaum), adam.seegmiller@vanderbilt.edu (A.C. Seegmiller), charles.stratton@vanderbilt.edu (C. Stratton), payton.j.odaniels@vanderbilt.edu (P. O'Daniels), william.holmes@vanderbilt.edu (W.R. Holmes).
[1] Authors contributed equally.

https://doi.org/10.1016/j.cognition.2021.104713
Received 22 September 2020; Received in revised form 1 March 2021; Accepted 27 March 2021
Available online 3 April 2021

prevalence effect.

An alternative theory of decision-making that is able to disentangle different types of biases is the Diffusion Decision Model (DDM), a popular tool for studying speeded decision tasks (Ratcliff, 1978; Ratcliff, Smith, Brown, & McKoon, 2016) and more recently linked to the neurobiology of decision-making (Ratcliff & Smith, 2004). Importantly the DDM can distinguish between a "response expectancy bias" and a "stimulus evaluation bias" (Leite & Ratcliff, 2011; Ratcliff, 1985; Ratcliff, Van Zandt, & McKoon, 1999; White & Poldrack, 2014). It is able to untangle these latent biases by using individuals' joint choice and response time distributions (rather than simply choices as in SDT). The DDM is part of a class of models known as Evidence Accumulation Models (EAMs), which assume that perceptual decisions are made through the accumulation of evidence for different alternatives over time. Recent work has shown EAMs provide an important link between the time-course of behavioral decisions and neural firing data (Ratcliff & Smith, 2004). In the model, a response expectancy bias corresponds to a shift in the starting point of evidence accumulation. A stimulus evaluation bias corresponds to a shift in the mean rate of evidence accumulation (the drift rate). Fig. 1 shows the results of a simulation study where SDT was fit to two simulated data sets from the DDM, one with a response expectancy bias and one with a stimulus evaluation bias. Results show that the response expectancy and stimulus evaluation biases influence the criterion parameter in SDT. Neither impacts discriminability. Thus, the DDM can distinguish between these distinct biases while SDT cannot. Simulation details are in the supplement.

Here we apply the DDM to data from two experiments where participants decided whether blood cell images are cancer related or not. The classification of peripheral blood cells is an important step in the diagnosis of malignant blood diseases such as leukemia and lymphoma. This task is often performed by medical technologists, who classify hundreds of cells from each patient. Automated systems are often used to aid medical technologists by pre-classifying images, which the technologists verify and reclassify when necessary. Experiment 1 involves novice participants, with no prior experience while Experiment 2 tests medical experts. In both, we manipulate the prevalence rate of cancerous cells. Results of the DDM suggest that prevalence influences both response expectancy and stimulus evaluation biases, with novices showing a more pronounced response expectancy bias and experts showing a more pronounced stimulus evaluation bias. All of the data are available at https://osf.io/4n7sr/.

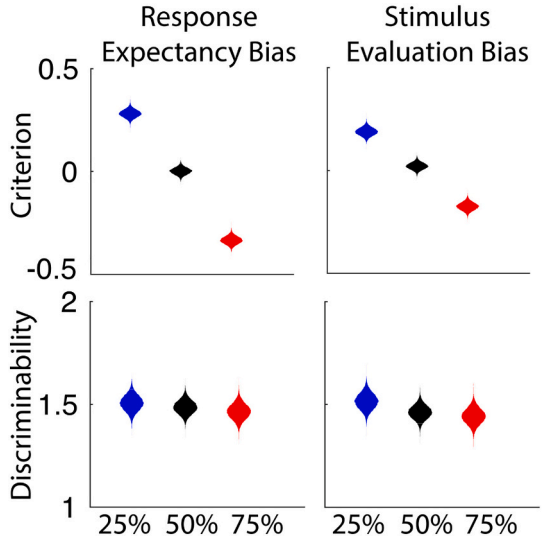

Fig. 1. DDM and STD simulation study - Results of a simulation study where SDT was fit to two simulated data sets from the DDM, one with a response expectancy bias (left panels) and one with a stimulus evaluation bias (right panels). Each simulated data set contained three prevalence conditions: low prevalence (25% targets), equal prevalence (50% targets), and high prevalence (75% targets). SDT was fit separately to each prevalence condition in the two simulated data sets using Bayesian methods. The violin plots show the posterior distributions for the criterion and discriminability parameters.

2. Experiment 1: novice participants

Experiment 1 tests the prevalence effect with novice participants, providing an important baseline for evaluating experts (Experiment 2). Further, all medical image observers are novices prior to training. Thus, studying novices provides valuable information on potential ways to augment training in the future.

2.1. Behavioral methods

2.1.1. Participants

39 Vanderbilt University undergraduate students (age range = [18, 28], mean = 19.7; 77% female) participated in Experiment 1a and 57 students (age range = [18, 22], mean = 19.2; 58% female) participated in Experiment 1b, both for course credit. The sample size was selected based on modeling requirements. The typical sample size for experiments using similar modeling methods is 20–40 participants per condition (Dutilh et al., 2012; Holmes, Trueblood, & Heathcote, 2016; White & Poldrack, 2014). Experiment 1a used a within-subjects design and Experiment 2a had two between-subject conditions. The experimental protocol was approved by the institutional review board at Vanderbilt University.

2.1.2. Materials

The stimuli were 300 digital images of Wright-stained white blood cells taken from anonymized patient peripheral blood smears at Vanderbilt University Medical Center (VUMC) used in Trueblood et al. (2018), which examined speed / accuracy tradeoffs in medical image decision-making.[2] The images were taken by an automated digital cell morphology instrument called the CellaVision DM96 (CellaVision AB, Lund, Sweden). Half of the images contained blast cells and the other half contained non-blast cells (see Fig. 2 for examples). In each category, half of the images were "easy" and half were "hard". Thus, in total, there were 75 images in each of the four following categories: easy blast, hard blast, easy non-blast, and hard non-blast. These classifications were based on ratings from three hematopathology faculty from the Department of Pathology at VUMC. The raters first identified each image as a blast or non-blast cell and then provided a difficulty rating for the image on a 1–5 scale. Details of the rating procedure and classification process can be found in Trueblood et al. (2018).

2.1.3. Procedure

In both Experiments 1a and 1b, participants first read instructions explaining that the images were of white blood cells and that their task was to decide whether or not each image contained a blast cell. Participants were told that a blast cell is a pathological white blood cell whose presence is often a sign of blood cancers such as leukemia. Following the initial instructions, participants completed a learning phase where they were shown a single image with its category label, either 'Blast' or 'Non-blast'. Each trial started with a fixation cross displayed for 200 milliseconds. Following the fixation cross, the image was displayed in the center of the screen. Participants were allowed to examine the image for as long as they liked. When they were finished with a particular image, they pressed the space bar to see the next one. They viewed 36 randomized images (9 examples from each of the four image categories). Next, the participants completed a training phase where they were

[2] While the images are the same as in Trueblood et al. (2018), this paper examines a different phenomenon (i.e., the prevalence effect) and reports new data.

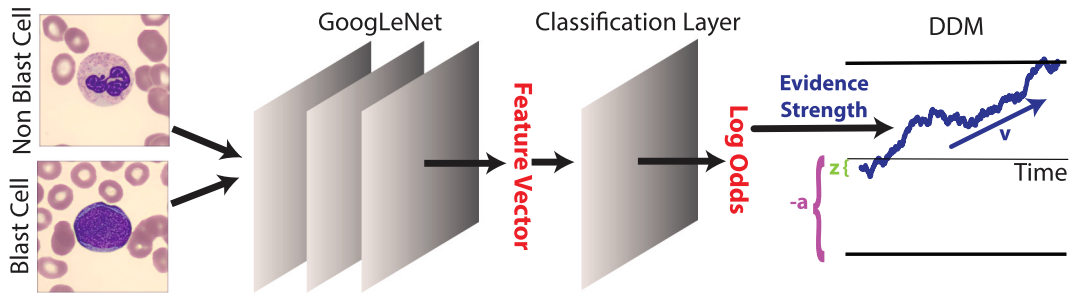

Fig. 2. Modeling overview – A convolutional neural network (CNN) is coupled with the diffusion decision model (DDM) to model decisions at the level of individual images. The two images on the left are representative of images used in this task. A pre-trained version of GoogLeNet was used as the base for our CNN. A new classification layer was added and the final 10 layers of the resulting network were trained using transfer learning on the medical image data set. The resulting CNN assigns to each image a probability of being either a Blast or Non-Blast cell. This probability is then used to construct an image specific drift rate that is input into the DDM.

shown two images (one blast image and one non-blast image) and a single label, either 'Blast' or 'Non-blast'. The participants were instructed to select the image that matched the category label. They completed 60 randomized training trials comprised of 15 trials from the following image pairings: easy blast with easy non-blast, easy blast with hard non-blast, hard blast with easy non-blast, and hard blast with hard non-blast. Participants received trial-by-trial feedback (displayed for 500 ms after each trial). Following the training phase, the procedures for Experiments 1a and 1b diverged.

In Experiment 1a, participants completed three blocks of practice trials similar to the main task. On each trial, participants were shown a single image and were asked to decide if it was a blast or non-blast cell. Each trial started with a fixation cross displayed for 200 milliseconds. Following the fixation cross, the image was displayed in the center of the screen and participants were given up to 5 s to respond. The image was displayed during the entire response period (up to 5 s), and a response terminated the trial. Participants received trial-by-trial feedback (displayed for 500 ms after each trial). Each practice block contained 48 randomized trials. In the high prevalence block, 75% of the images contained a blast cell (i.e., 36 blast trials comprised of 18 hard and 18 easy images) and 25% of the images contained a non-blast cell (i.e., 12 non-blast trials comprised of 6 hard and 6 easy images). In the equal prevalence block, 50% of the images contained a blast cell (equal numbers of easy and hard trials) and 50% of the images contained a non-blast cell (equal numbers of easy and hard trials). In the low prevalence block, 25% of the images contained a blast cell (i.e., 12 blast trials comprised of 6 hard and 6 easy images) and 75% of the images contained a non-blast cell (i.e., 36 non-blast trials comprised of 18 hard and 18 easy images). At the beginning of each block, participants were told the proportion of blast and non-blast cells for that block. The order of the three blocks were randomized across participants.

After participants completed the three practice blocks, they started the main portion of the experiment. There were three block types: high (75% blast), equal (50% blast), and low (25% blast) prevalence. Each block contained 48 randomized trials and the structure of the blocks was identical to the practice. Participants were told the proportion of blast and non-blast cells at the beginning of each block. Each type of block was repeated 7 times for a total of 21 blocks. The order of the blocks was fully randomized. Participants did not receive feedback.

In Experiment 1b, participants were randomly assigned to one of two groups following the training phase: the high prevalence group or the low prevalence group. All participants completed one practice block at equal prevalence (50% blast and 50% non-blast images). The block contained 80 randomized trials with an equal number of hard and easy trials (i.e., 20 easy blast, 20 hard blast, 20 easy non-blast, and 20 hard non-blast trials). Participants were told the proportion of blast and non-blast cells at the beginning of the block. On each trial, participants were shown a single image and were asked to decide if it was a blast or non-blast cell. Each trial started with a fixation cross displayed for 200 milliseconds. Following the fixation cross, the image was displayed in the center of the screen and participants were given up to 5 s to respond. Participants received trial-by-trial feedback (displayed for 500 ms after each trial).

Following the practice block, participants started the main portion of the experiment. Both groups first completed two blocks of 80 randomized trials at equal prevalence. The structure of these blocks was identical to the practice. After the two equal prevalence blocks, the high prevalence group completed 12 blocks at 90% prevalence. Each block contained 80 randomized trials (36 easy blast, 36 hard blast, 4 easy non-blast, and 4 hard non-blast trials). The low prevalence group completed 12 blocks at 10% prevalence. Each block contained 80 randomized trials (4 easy blast, 4 hard blast, 36 easy non-blast, and 36 hard non-blast trials). Participants were told the proportion of blast and non-blast cells at the beginning of each block. Participants did not receive feedback.

The procedural differences between Experiments 1a and 1b were due to the more extreme prevalence rates used in Experiments 1b. Specifically, when running experiments with extreme prevalence rates, a large number of trials are needed in order to have a sufficient number of low-prevalence trials for data analysis and modeling. In order to keep the length of Experiment 1b under one-hour, we used a between-subjects design and reduced the number of equal prevalence trials.

2.2. Behavioral results

No participants were excluded from the analyses. The mean accuracy on the task was M = 0.695 (SD = 0.091) for Experiment 1a. A Bayesian one sample *t*-test provided strong evidence that this value was greater than chance ($BF_{+0} = 2.361e{+}13$) with a 95% credible interval of [0.666, 0.724].[3] For Experiment 1b, mean accuracy on the task was M = 0.754 (SD = 0.117). Similar to Experiment 1a, a Bayesian one sample *t*-test showed strong evidence that accuracy was greater than chance ($BF_{+0} = 1.701e{+}20$) with a 95% credible interval equal to [0.723, 0.785]. Thus, we conclude that novice participants learned to distinguish between blast and non-blast cells. For the main trials, we analyzed the choice data using a Bayesian logistic mixed-effects regression model in order to test (1) whether a prevalence effect occurred and (2) to examine whether the strength of the prevalence effect depended on the time at which a decision was made. As illustrated in White and Poldrack (2014), response expectancy and stimulus evaluation biases affect different parts of the RT distribution. Thus, if we observe that the prevalence effect depends on decision time, it suggests that these biases might be present in our data and motivates our approach of jointly modeling choice and

[3] BF_{+0} is the Bayes Factor for the one-sided *t*-test testing the alternative hypothesis that the mean is greater than 0.5. Tests were conducted in JASP (JASP Team, 2020).

response time data with the DDM. In our model, response (coded as 1 = blast and 0 = non-blast) was the dichotomous dependent variable. The fixed effects included prevalence rate (with reference level set to 50% prevalence), difficulty (coded as 1 = hard, and 0 = easy), image type (coded as 1 = blast and 0 = non-blast), and RT quantile calculated on an individual basis by using nine evenly spaced cut points resulting in 10 RT bins. We also included the interaction between RT quantile and prevalence rate as this was one of our main questions of interest. No other interaction terms were included. We also allowed for by-subject random intercepts.[4] The model was fit using JASP (JASP Team, 2020).

We analyzed Experiments 1a and 1b separately. For Experiment 1a, we ran three MCMC chains for 4000 iterations with a burn-in of 2000 iterations. The model converged with all R-hat values less than 1.01. Table 1 shows the estimated marginal means for each prevalence condition for the 10 RT quantiles (the fixed effects parameter estimates and 95% credible intervals are provided in the supplement). As shown in Table 1, the prevalence effect was the largest for fast response times, and there was less difference in the probability of responding blast for the three prevalence rates when responses were slow. Specifically, for 25% prevalence, the estimated marginal means for the first four RT quantiles were less than 0.50 (i.e., the 95% Highest Posterior Density (HPD) interval was less than 0.50), showing that participants responded non-blast more often when they made quick responses. Likewise, for 75% prevalence, the estimated marginal means for the first four RT quantiles were greater than 0.50 (i.e., the 95% HPD interval was greater than 0.50), showing that participants responded blast more often when they made quick decisions. Fig. 3 further illustrates the relationship between choice, prevalence, and response time by showing that the largest proportion of false alarms occurred at 75% prevalence for short responses and the largest proportion of misses occurred at 25% prevalence for short responses.

Table 1
Estimated marginal means for Experiment 1a.

RT quantile	Prevalence	Median	95% HPD	
			Lower	Upper
1	25	**0.197**	0.164	0.233
2	25	**0.299**	0.256	0.343
3	25	**0.372**	0.324	0.419
4	25	**0.411**	0.365	0.460
5	25	0.452	0.406	0.503
6	25	0.485	0.437	0.535
7	25	0.483	0.436	0.533
8	25	0.530	0.481	0.578
9	25	0.522	0.474	0.572
10	25	0.525	0.474	0.572
1	50	0.448	0.399	0.500
2	50	0.468	0.422	0.520
3	50	0.481	0.429	0.526
4	50	0.506	0.457	0.554
5	50	0.478	0.429	0.526
6	50	0.504	0.456	0.552
7	50	0.535	0.490	0.586
8	50	0.519	0.470	0.567
9	50	0.478	0.431	0.526
10	50	0.479	0.431	0.527
1	75	**0.697**	0.653	0.739
2	75	**0.624**	0.580	0.674
3	75	**0.603**	0.554	0.648
4	75	**0.553**	0.506	0.603
5	75	0.533	0.483	0.581
6	75	0.559	0.510	0.607
7	75	0.491	0.444	0.543
8	75	0.503	0.459	0.556
9	75	0.463	0.415	0.510
10	75	0.462	0.416	0.513

Note. Results are averaged over the levels of image type and difficulty.
Note. Results are on the response scale.
Note. Estimates in bold have 95% HPD intervals that exclude .50.

To further test the interaction between prevalence and RT quantile, we compared a model without the interaction term to the model including this term. The Widely Applicable Information Criterion (WAIC) for the model without the interaction term was 46,324.58. Including the interaction term in the model reduced the WAIC to 45,695.20 (smaller is better), thus showing the interaction term improves fit despite the additional complexity.

The results for Experiment 1b were similar to those of Experiment 1a. For this experiment, we ran 3 MCMC chains for 8000 iterations with a burn-in of 2000 iterations. The model converged with all R-hat values less than 1.01. Table 2 shows the estimated marginal means for each prevalence condition for the 10 RT quantiles (the fixed effects parameter estimates are provided in the supplement). As shown in Table 2, for 10% prevalence, the estimated marginal means for the first seven RT quantiles were less than 0.50. Likewise, for 90% prevalence, the estimated marginal means for the first seven RT quantiles were greater than 0.50. Thus, the prevalence effect was observed for quick responses.

Similar to Experiment 1a, we also compared a model without the interaction term to the model including this term. The WAIC for the model without the interaction term was 63,155.52. Including the interaction term in the model reduced the WAIC to 60,202.38, thus showing the interaction term improves fit despite the additional complexity.

In addition to fitting Bayesian logistic mixed-effects regression models to the choice data, we also fit an equal-variance SDT model using hierarchical Bayesian methods following Lee and Wagenmakers (2014) using Matlab and JAGS (Plummer, 2003). We ran three MCMC chains, set the burn-in to 1000 samples, and recorded 10,000 samples after the burn-in. Fig. 3 (right panels) shows the posterior distributions for the hierarchical criterion and discriminability parameters for Experiments 1a (top) and 1b (middle). As seen in the figure, prevalence had a large impact on criterion and little to no impact on discriminability. One of the benefits of using Bayesian parameter estimation is that we can perform significance testing directly on posteriors (Kruschke, 2010). In the figure, we include the proportion of the posteriors in which the parameter for one condition (e.g., criterion at low prevalence) is greater than the parameter for another condition (e.g., criterion at equal prevalence). For the criterion parameter, these proportions are close to or equal to 1, implying there is very little to no overlap in the posteriors.

2.3. CNN-DDM modeling methods

In our task, each trial uses a distinct image. Thus, rather than combine images into "like" classes that are grouped for modeling (which is most common), we will model each image separately. This requires incorporating image specific information into the decision model. More specifically, it requires accounting for the fact that there is a continuum of difficulties associated with these images that needs to be reflected in the parameterization of drift rates in the DDM. To incorporate this image specific information into our modeling, we couple a deep convolutional neural network (CNN) with the DDM to model choices and response times (Holmes, O'Daniels, & Trueblood, 2020). In this model, the CNN is used to extract image specific information (image type and strength of information) that is then incorporated into the DDM.

At a broad level, this model uses a partially custom CNN to assign a probability of being a blast cell (denoted by P_{Blast}) to each image in the task, which is then transformed into a drift rate to represent that image in the DDM. In this way, the CNN extracts information directly from each image, and passes that information to the decision model. This joint CNN-DDM is then fit, at the level of individual trials, to the choice and response time (RT) data using hierarchal Bayesian techniques. See Fig. 2

[4] By-subject random slopes were not included because of convergence issues with the addition of these parameters.

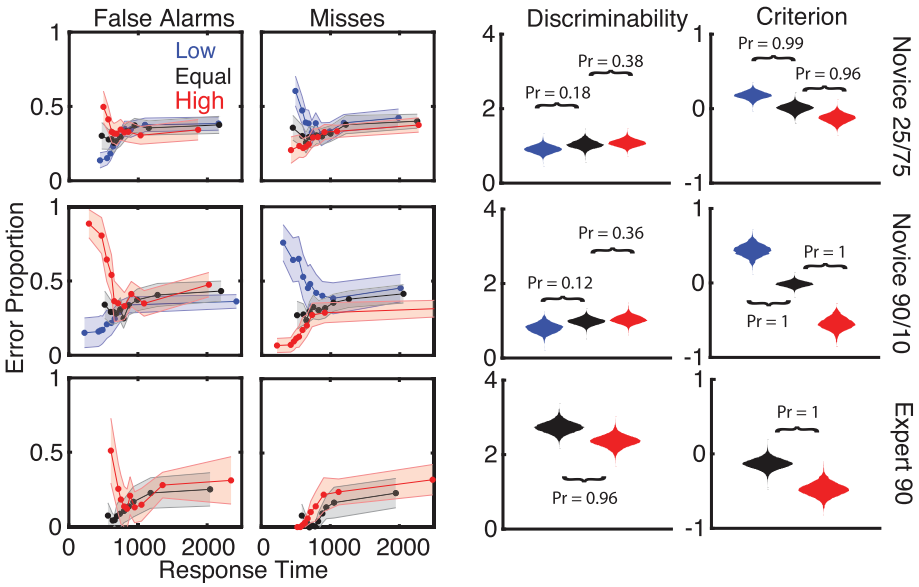

Fig. 3. Behavioral and SDT results – Left panels show the proportion of errors for novices (top two rows) and experts (bottom row) as a function of response time and prevalence. The mean error proportion and 95% confidence intervals (shaded error bands) were calculated for responses that fell within time bins defined by 10 quantiles of the response time distribution, calculated on an individual basis. The median of each RT quantile is plotted on the x-axis. Right panels show the SDT modeling results. Violin plots show the posterior distributions for the hierarchical criterion and discriminability parameters where black indicates 50% prevalence while blue and red indicate low and high blast prevalence, respectively. The numbers overlaid on each plot quantify the proportion of the posterior samples where 1) the low prevalence condition is greater than that of the 50% condition, 2) the high prevalence condition is less than that of the 50% condition, or 3) the low prevalence condition is greater than the high prevalence condition. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Estimated marginal means for Experiment 1b.

RT quantile	Prevalence	Median	95% HPD	
			Lower	Upper
1	10	**0.202**	0.166	0.239
2	10	**0.225**	0.187	0.266
3	10	**0.236**	0.196	0.278
4	10	**0.270**	0.227	0.315
5	10	**0.306**	0.260	0.354
6	10	**0.357**	0.307	0.408
7	10	**0.405**	0.354	0.459
8	10	0.459	0.406	0.516
9	10	0.493	0.438	0.547
10	10	0.500	0.446	0.556
1	50	0.485	0.382	0.586
2	50	0.538	0.454	0.625
3	50	**0.583**	0.510	0.658
4	50	0.513	0.443	0.579
5	50	0.557	0.494	0.621
6	50	0.513	0.450	0.572
7	50	0.498	0.440	0.554
8	50	0.480	0.424	0.535
9	50	0.508	0.453	0.562
10	50	0.490	0.436	0.545
1	90	**0.905**	0.882	0.925
2	90	**0.894**	0.870	0.916
3	90	**0.841**	0.808	0.871
4	90	**0.819**	0.784	0.852
5	90	**0.762**	0.720	0.802
6	90	**0.678**	0.630	0.726
7	90	**0.602**	0.548	0.654
8	90	0.546	0.491	0.600
9	90	0.541	0.487	0.597
10	90	0.538	0.484	0.593

Note. Results are averaged over the levels of image type and difficulty.
Note. Results are on the response scale.
Note. Estimates in bold have 95% HPD intervals that exclude 0.50.

for an overall schematic of this joint modeling approach (see Holmes, O'Daniels, & Trueblood, 2020, where this model was originally developed, for further details).

The coupling of a CNN to extract image information with the DDM has several benefits. First, it allows us to extract unique information about each individual image, rather than group them into classes as is most common. We will evaluate the benefit of using image specific information by comparing the CNN-DDM model to a standard DDM model where the images are grouped into four categories (Easy, Hard x Blast, Non-Blast). Second, it substantially lowers the burden of curating an image set like this. Expert classification of only enough images to train the CNN are needed and subsequently more images can be added without extra input. Additionally, only class information about those initial images is required, which is simpler to obtain than more complex information such as classification difficulty.

2.3.1. *Diffusion decision model*

The CNN-DDM incorporates a version of the standard diffusion decision model in order to account for choices and response times (Ratcliff, 1978). Specifically, we use a version that includes non-decision time (t_{ND}), start point bias (z), threshold (a), and a stimulus dependent drift rate (v). In the CNN-DDM model, the drift rate (v_i) for image "i" is derived from the characteristics of the medical image itself using the CNN. In this way, the characteristics of individual images, as determined by the CNN, determine the strength of evidence that forms the basis of the drift rate for that individual image. In the subsequent sections, we discuss how the CNN outputs are incorporated into the drift rates and discuss the drift rate parameters that will be estimated. For simplicity and to make hierarchal Bayesian fitting of this data tractable, we do not include trial-to-trial noise variations in either start point or drift rate (typically referred to as s_z and s_v). In the CNN-DDM model, each image is treated as a single trial condition, and thus it is highly unlikely that these parameters would be estimable.

2.3.2. *Translating images to probabilities using a deep CNN*

To translate each medical image into a single numeric probability of being a Blast or non-Blast, we augmented a GoogLeNet deep CNN (Szegedy et al., 2015) that was pre-trained on the ImageNet database (we downloaded the fully pre-trained network). What follows is modestly technical and requires some knowledge of neural networks and their training. In short though, we partially re-trained this network using our cell images so that, rather than discriminating between ImageNet categories (e.g. identifying cats), it can discriminate between blast and non-blast cells. The output of this CNN will be a probability that describes how confident the network is that a particular image is a Blast cell. This is then the basis of forming an image dependent drift rate for the DDM component of the CNN-DDM model.

We took the pre-trained GoogLeNet network and removed the classification layer, leaving only the layers that translate an image into a feature vector (FV). We then added a single softmax classification layer

that translates that FV into a probability of being a Blast cell (P_{Blast}). In neural networks, a softmax classification layer transforms the numeric output of the last neuron layer of a neural network into probabilities that sum to one. It does this by taking the exponents of each output and then normalizing each number by the sum of those exponents. We chose a softmax layer specifically since its output can be interpreted as a probability of belonging to a particular class (i.e., the probability of the image being a blast cell). Transfer learning was then used to train this network on an image bank consisting of 606 images (326 Blast and 280 NonBlast), of which 80% were used for training and 20% for validation. Transfer learning takes a pre-trained network, trained on a dataset of similar type, and refines the network's parameters to perform on the new dataset. In our case, the network being used is far too large to be trained from scratch on our relatively small cell image dataset. Thus, we utilized the pre-trained version, trained on the ImageNet dataset, and fine-tuned its training on our cell image data. This is a common procedure in deep learning to extend a network from one type of data (e.g. cats) to another (Blast cells, in our case).

After training of this network, it had a classification accuracy of 94% on the validation set and 98% on the training set. For further details about the training, see Holmes, O'Daniels, & Trueblood (2020). A Matlab Live Script (Matlab's equivalent of a Jupyter Notebook) that creates this trained network can also be found on the Open Science Framework page associated with that paper.

At completion of training, the resulting network is used to extract classification probabilities for each image. This is then transformed into a log-odds value via $LO = \log(P_{Blast}/P_{NonBlast})$. The logarithm of the odds ratio (i.e. log-odds) is a useful transformation that symmetrizes the resulting probabilities (LO is positive and negative where probabilities are positive only). For an individual image, the output log-odds value is then used to construct a drift rate as described in the next section. Note that a positive (resp. negative) *LO* value indicates that the image is a Blast (resp. Non-Blast) cell and larger magnitudes of *LO* indicate a higher confidence in that classification.

2.3.3. Coupling the CNN with the diffusion model

The CNN described above takes each medical image and produces a measure of the stimulus information (i.e., the log odds) for that image. To connect the output of the CNN (*LO*) to the input of the DDM (drift rate), we utilize a linear function to map *LO* values into a drift rate $v = v_{prev} + v_{slope} * LO$. Here, v_{prev} and v_{slope} are participant level model parameters that are estimated. v_{slope} (i.e. the drift slope) measures how effective a participant is at integrating information from an image; large values indicate high effectiveness and small values low effectiveness. In the extreme, $v_{slope} = 0$ implies that an individual does not incorporate any image specific information into their decision process. For example, this might occur for someone entirely unfamiliar with white blood cells. In fitting this model to our data, a single value of this parameter is used, independent of the Blast prevalence used in the particular experimental condition. This parameter is intended to measure the prevalence independent component of image information extraction. Note that while we utilize a linear mapping from log-odds to drift rate in this application, in principle more complex mappings could potentially be included. For example, a saturating function that asymptotes at large log-odds values could be incorporated. There is no technical restriction on this functional from an implementation perspective. However, it does lead to more parameters to estimate and a generally more complex model. Since the questions at hand do not warrant this added complexity, we have opted for simplicity in this application.

The v_{prev} parameter (i.e. the drift intercept) measures how strongly prevalence influences people's evaluation of the stimulus. The sign of this parameter encodes whether prevalence introduces a bias favoring the perception of blast (positive sign) or non-blast characteristics (negative sign). Similarly, the magnitude describes how large or small that effect is. A small value (relative to the $v_{slope} * LO$) term indicates little effect of prevalence while a larger value indicates a more substantial bias. In the extreme, if $v_{prev} = 0$, then there would be no influence of prevalence on the evidence accumulation process. This parameter is independent of the image itself. However, different values are independently estimated for each prevalence condition.

2.3.4. Accounting for prevalence with model parameters

The purpose of this study is to assess how prevalence influences people's decisions. We consider two possible biases that prevalence might introduce: stimulus evaluation bias and response expectancy bias (Leite & Ratcliff, 2011; White & Poldrack, 2014). The response bias accounts for the possibility that high or low prevalence rates may simply cause people to have a stimulus independent, pre-disposition to choose either blast or non-blast, reflecting a strategy to choose the more likely response. This bias is encoded in the start point parameter, which captures stimulus independent initial biases. Thus for each experimental prevalence condition, we estimate a separate response bias parameter. In Experiment 1a, three bias parameters will be estimated (z_{25}, z_{50}, z_{75}) corresponding to the 25, 50, and 75% conditions respectively. For Experiment 1b, two bias parameters will be estimated for each between-subjects condition (z_{10}, z_{50} in the 10/50 condition and z_{50}, z_{90} in the 50/90 condition).

The second type of bias is the stimulus evaluation bias. This bias accounts for the possibility that high or low prevalence may alter how people evaluate the image. This is accounted for in the previously described v_{prev} parameter. In Experiment 1a, three parameters will be estimated (v_{25}, v_{50}, v_{75}) corresponding to the 25, 50, and 75% conditions respectively. For Experiment 1b, two stimulus bias parameters will be estimated for each between-subjects condition, (v_{10}, v_{50} in the 10/50 condition and v_{50}, v_{90} in the 50/90 condition).

Since we will be drawing inferences about the values of fit parameters in this model, particularly its bias related parameters, we perform parameter recovery to ensure the model is identifiable and parameters are recoverable. To do so, we simulated data out of a synthetic version of Experiment 2. That is, we simulated synthetic data with the same task structure and amount of data found in our experiments. We then used Bayesian parameter estimation to fit this model to that data. Results (SM Fig. 5) demonstrate that the model's parameters are precisely estimable with the type of data coming from this task.

2.3.5. Hierarchical Bayesian parameter estimation

We fit the CNN-DDM to choice-RT data in order to assess the effect of prevalence on people's decision processes. As described in the Experimental Methods section, each experiment consists of different prevalence conditions. For each experiment, a single value of the threshold (a), non-decision time (t_{ND}), and drift slope (v_{slope}) are estimated since these components of the decision process are not posited to depend on image prevalence. In Experiment 1a (resp. 1b), there are also an additional six (resp. four) bias parameters that will be estimated to measure the effect of prevalence on decisions.

For each experiment, all prevalence conditions are simultaneously fit to data using Hierarchal Bayesian parameter estimation. Using this approach, we estimate both group and individual level parameters simultaneously. All figures depict posterior distributions for hyper mean parameters. The DEMCMC algorithm (Turner & Sederberg, 2012) was used. In order to efficiently calculate the DDM likelihood function, we used a variant of the algorithm in Navarro and Fuss (2009) where the short and long time series expansion times of the WFPT (Weiner First Passage Time) infinite series was truncated at four terms (which yield errors <1e-5). This approach is much more effective than other approaches such as the Probability Density Approximation (Evans, Holmes, & Trueblood, 2019; Evans, Trueblood, & Holmes, 2019; Holmes, 2015; Holmes et al., 2016; Holmes & Trueblood, 2017; Lin, Heathcote, & Holmes, 2019; Trueblood, Heathcote, Evans, & Holmes, 2021; Turner & Sederberg, 2014), though is only applicable to this precise form of the DDM. The intra-trial variability parameter for the DDM was fixed at $s = 0.1$ as is common. See the Supplemental Material

for specification of the model priors.

2.4. Modeling results

To assess the influence of prevalence on decisions in Experiments 1a and 1b, we compared the posterior distributions for the start point (z) and drift (v_{prev}) parameters across different prevalence conditions (Fig. 4, top two rows). Posterior distributions for all other parameters are shown in SM Fig. 1. In Fig. 4, posterior distributions are calculated as the difference between an extreme prevalence parameter and the corresponding 50% prevalence parameter (e.g., z_{25} - z_{50}). By plotting the difference from 50%, it is easy to visualize changes in parameter values for extreme prevalence conditions as compared to equal prevalence. For example, the top left plot compares the response expectancy bias between the 25 and 50% (blue) conditions as well as 50 and 75% conditions (red, note that the posterior is centered on the horizontal line at zero). For each difference, we directly calculate the probability (stated as $p = 0.97$, for example, on these plots) that the two posteriors are different (similar to the SDT analysis). Bayesian 95% credible intervals along with posterior mean values for these parameter differences are provided in Table 3.

For purposes of plotting in Fig. 4, we have normalized the response bias and stimulus bias parameters. The vertical axis of the "response bias" (e.g. start point) plots are measured as a percentage of the response threshold. This effectively normalizes the scale of the prevalence effect in the response bias parameter and facilitates comparison across experiments. Stimulus bias parameters are also normalized for presentation purposes. Recall that $v = v_{prev} + v_{slope} * LO$ in the model. The nominal log-odds absolute value from the CNN is ~10. For the stimulus bias figures, we normalized the prevalence parameter according to v_{prev}/(v_{slope} * 10) before calculating the relevant differences for plotting. The vertical scale on these plots thus measures the strength of the prevalence effect on the drift rate relative to the prevalence independent drift component. Thus, a value of *0.1* on this scale indicates that prevalence causes a roughly 10% shift in the drift rate relative to the log-odds component. This normalization facilitates comparison of stimulus biases across experiments.

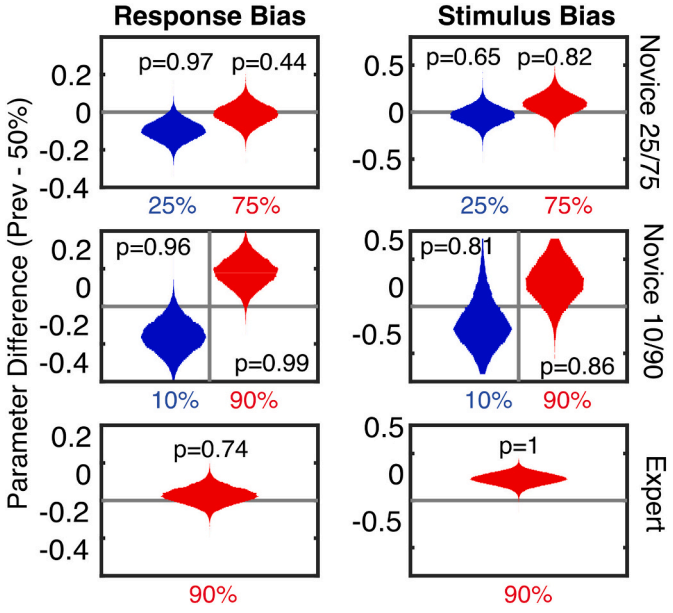

Fig. 4. Quantifying the effect of prevalence on response and stimulus biases – Each row represents the results of Experiments 1a, 1b, and 2 respectively. The first and second columns illustrate the effect of prevalence on the response expectancy bias and stimulus evaluation bias, respectively. Each plot shows the posterior distribution for the relevant parameter difference. For example, in the top left plot the 25% (blue) distribution indicates the posterior distribution of the difference between the 25% and 50% response bias parameters. The numbers overlaid on each plot quantify the proportion of the posterior samples where either 1) the low prevalence bias is less than that of the 50% condition or 2) the high prevalence bias is greater than that of the 50% condition. Posterior mean values and 95% credible intervals are provided in Table 3. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Bayesian 95% credible intervals (CI) and posterior mean values for parameter differences plotted in Fig. 4. In each case, the deviation of the relevant extreme prevalence (e.g. 25, 75, 10, 90%) parameter from the companion 50% parameter is analyzed, as noted in each row. As an example, for the first row, the 95% credible interval for the posterior distribution of z_{25}-z_{50} is determined so it can be assessed whether the full CI lies below 0. Intervals excluding 0 are in bold. Posterior mean values of these differences are also provided. Bayes *p*-values for these differences are provided on Fig. 4.

Parameter	Experiment	95% Credible Interval	Posterior Mean
Response expectancy bias	Exp 1a, (z_{25} – z_{50})	**−0.199, −0.0039**	−0.097
	Exp 1a, (z_{75} – z_{50})	−0.114, 0.095	−0.009
	Exp 1b, (z_{10} – z_{50})	**−0.303, 0.037**	−0.134
	Exp 1b, (z_{90} – z_{50})	**0.032, 0.327**	0.182
	Exp 2, (z_{90} – z_{50})	−0.057, 0.114	0.027
Stimulus evaluation bias	Exp 1a, (v_{25} – v_{50})	−0.216, 0.142	−0.034
	Exp 1a, (v_{75} – v_{50})	−0.102, 0.302	0.091
	Exp 1b, (v_{10} – v_{50})	−0.735, 0.468	−0.233
	Exp 1b, (v_{90} – v_{50})	−0.180, 0.826	0.297
	Exp 2, (v_{90} – v_{50})	**0.104, 0.347**	0.230

Analysis of the Novice 25/75% (Experiment 1a) results shows that low prevalence (25%) introduces a substantial (and significant) response expectancy bias but does not appear to introduce any stimulus evaluation bias. Comparing the 50/75% conditions shows no apparent effect on the response expectancy bias, but there appears to be a weak stimulus evaluation bias. Analysis of the Novice 10/90% (Experiment 1b) data shows that in both high and low prevalence conditions, there appears to be a significant response expectancy bias. We also observe a stimulus evaluation bias, although this effect appears weaker than the effect on the response expectancy bias. As shown in Table 3, the only parameter differences with 95% CIs that exclude 0 are those involving the response expectancy bias parameters (highlighted using bold font in the table). We thus conclude that high or low prevalence conditions introduce a strong response expectancy bias and a weak stimulus evaluation bias in novice populations.

To assess the quality of fit, we compared the CNN-DDM model described above to a baseline version of the model with only a single start point and drift intercept (that is, these parameters do not vary with prevalence). This approach follows the recommendations of Lee et al. (2019) to use "bookend models" to assess quality of fit. This approach quantifies whether the addition of prevalence related parameters improves the fit to data. Results in Fig. 5 indicate that for the significant majority of participants, including different response and stimulus bias parameters for each prevalence condition improves the quality of fit to the data. The same approach will be used to assess quality of fit in subsequent experiments.

We also fit an intermediate model to the data that is more complex than this bookend model but does not use the CNN to construct the drift rate. In this model, the drift rate is again constructed as combination of prevalence (v_{prev}) and image dependent (v_{class}) components ($v = v_{prev} + v_{class}$). However, in this case, rather than considering each image

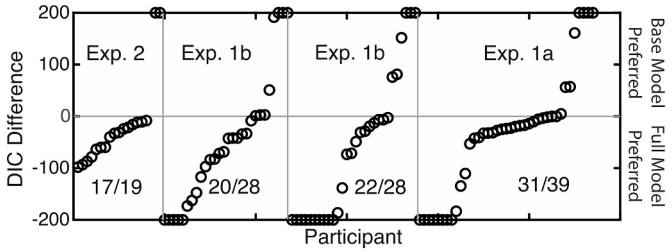

Fig. 5. Quality of model fit: For Experiments 1a, 1b, and 2, the Deviance Information Criterion (DIC) for the full prevalence model is compared on a participant by participant basis to a baseline model that lacks prevalence effects. Each data point is the DIC difference for a single participant with negative values indicating the prevalence model is favored. Differences larger than 200 DIC points are truncated to 200 for visual purposes. All fractions in the lower quadrants indicate the fraction of participants where the prevalence model provides a superior fit.

individually, we group them into four categories (Easy, Hard x Blast, Non-Blast) according to expert difficulty ratings obtained during image curation (Trueblood et al., 2018). In this model, we estimate four v_{class} parameters, one for each image category. We refer to this as the discrete drift prevalence model (see SM for further discussion).

This model was fit to the data using the same procedures described previously. Parameter estimates show similar results to the CNN-DDM (SM Fig. 2 and SM Table 4). Further results demonstrate that the CNN-DDM model fits the data better than this discrete drift model for 75% of participants across all experiments (SM Fig. 3) and that the discrete drift prevalence model fits choice proportion and mean-RT data well (SM Fig. 4).

2.5. Conclusion

Experiment 1 showed the standard prevalence effect in novices: an increase in misses in low prevalence and an increase in false alarms in high prevalence. The DDM suggest these errors are due to a strong response expectancy bias and a weak stimulus evaluation bias.

3. Experiment 2: expert participants

Experiment 1 suggests that extreme prevalence leads to a strong response expectancy bias. Because novices lack experience with the task, they might simply adopt a strategy of selecting the more likely response (prevalence rates are clearly stated). This would naturally show up as a shift in the DDM start point. It is possible that well-trained experts would not utilize such as strategy. Experiment 2 replicates Experiment 1b (high prevalence condition only) with experts.

3.1. Behavioral methods

3.1.1. Participants

19 medical laboratory professionals (age range = [23, 61], mean = 38; 79% female) from Vanderbilt University Medical Center participated in the experiment online in exchange for a $15 gift card. The sample size for this experiment was based off of the availability of participants as well as modeling requirements. We note that this is a small size by typical standards for psychology experiments, but is very good for medical image perception where most studies have fewer than 20 participants. Due to the small number of medical experts available to participate, we only examined the high prevalence condition in this experiment. We decided to investigate high prevalence because it has traditionally been understudied in the medical image perception literature, and it is particularly relevant for pathology where initial filtering of images by automated systems results in humans seeing higher prevalence rates of abnormalities (even though prevalence of abnormalities is low in the general population). The experimental protocol was approved by the institutional review board at Vanderbilt University.

3.1.2. Materials

The stimuli were the same 300 digital images of Wright-stained white blood cells used in Experiment 1.

3.1.3. Procedure

Participants read the same initial instructions as in Experiment 1. Following the instructions, participants completed a training phase where they were shown two images (one blast image and one non-blast image) and a single label, either 'Blast' or 'Non-blast'. The participants were instructed to select the image that matched the category label. They completed 60 randomized training trials comprised of 15 trials from the following image pairings: easy blast with easy non-blast, easy blast with hard non-blast, hard blast with easy non-blast, and hard blast with hard non-blast. Participants received trial-by-trial feedback (displayed for 500 ms after each trial). Note that the participants did not complete the learning phase as in Experiment 1 since they were already familiar with the images. Following the training phase, the participants completed one practice block at equal prevalence (50% blast and 50% non-blast images). The block contained 40 randomized trials with an equal number of hard and easy trials (i.e., 10 easy blast, 10 hard blast, 10 easy non-blast, and 10 hard non-blast trials). Participants were told the proportion of blast and non-blast images at the beginning of the block. On each trial, participants were shown a single image and were asked to decide if it was a blast or non-blast cell. Each trial started with a fixation cross displayed for 200 milliseconds. Following the fixation cross, the image was displayed in the center of the screen and participants were given up to 5 s to respond. Participants received trial-by-trial feedback (displayed for 500 ms after each trial).

Following the practice block, participants started the main portion of the experiment. They first completed two blocks of 80 randomized trials at equal prevalence. The structure of these blocks was identical to the practice. After the two equal prevalence blocks, the participants completed 8 blocks at 90% prevalence. Each block contained 80 randomized trials (36 easy blast, 36 hard blast, and 8 hard non-blast trials). Note that these blocks did not contain easy, non-blast images. These images were excluded for several reasons. First, we excluded them in order to reduce the total experiment time to under 30 min, which is generally considered necessary when working with medical image observers. Second, based on results from Trueblood et al. (2018), medical experts have near ceiling performance on the easy, non-blast images. Third, the discrete drift model reported in the supplement requires a large number of observations per condition, which would have been difficult to achieve if both types of non-blast trials were included given the experiment time limitations. Participants were told the proportion of blast and non-blast cells at the beginning of each block. Participants did not receive feedback.

3.2. Behavioral results

No participants were excluded from the analyses. The mean accuracy on the task was M = 0.914 (SD = 0.042), higher than that for novices as expected. A Bayesian one sample *t*-test provided strong evidence that accuracy was higher than chance (BF_{+0} = 3.328 + 16) with a 95% credible interval equal to [0.894, 0.935]. For the main trials, we analyzed the choice data using a Bayesian logistic mixed-effects regression similar to Experiments 1a and 1b. As before, response (coded as 1 = blast and 0 = non-blast) was the dichotomous dependent variable. The fixed effects included prevalence rate (with reference level set to 50% prevalence), difficulty (coded as 1 = hard, and 0 = easy), image type (coded as 1 = blast and 0 = non-blast), and RT quantile calculated on an individual basis by using nine evenly spaced cut points resulting in 10 RT bins. We also included the interaction between RT quantile and prevalence rate as before. No other interaction terms were

included. We also allowed for by-subject random intercepts. The model was fit using JASP.

The results for Experiment 2, were similar to those of Experiment 1. For this experiment, we ran 3 MCMC chains for 4000 iterations with a burn-in of 2000 iterations. The model converged with all R-hat values less than 1.01. Table 4 shows the estimated marginal means for each prevalence condition for the 10 RT quantiles (the fixed effects parameter estimates are provided in the supplement). As shown in Table 4, for 90% prevalence, the estimated marginal means for the first seven RT quantiles are greater than 0.50 (also see Fig. 3). Thus, similar to the novices, experts showed a prevalence effect when they made quick decisions.

We also compared a model without the interaction between prevalence and RT quantile to the model including this term. The WAIC for the model without the interaction term was 7117.48. Including the interaction term in the model reduced the WAIC to 6762.76, thus showing the interaction term improves fit despite the additional complexity.

Similar to Experiments 1a and 1b, we also fit an equal-variance SDT model to the data. All modeling procedures were the same as before. The bottom, right panels in Fig. 3 show the posterior distributions for the hierarchical criterion and discriminability parameters for the expert participants. As seen in the figure, prevalence had a large impact on criterion. There is no overlap in the posteriors for the 50% condition and 90% condition (that is, $\Pr(50\% > 90\%) = 1$). Unlike the novice results, we do observe a difference in the posteriors for the discriminability parameter, with discriminability being a little higher for the 50% condition than the 90% condition. We note that the effect of prevalence on discriminability differs from previous findings showing that prevalence only effects criterion (Evans, Tambouret, Evered, Wilbur, & Wolfe, 2011; Horowitz, 2017; Wolfe et al., 2005; Wolfe et al., 2007; Wolfe & Van Wert, 2010).

3.3. Modeling methods

As with Experiment 1a, we fit the CNN-DDM to choice-RT data in order to assess the effect of prevalence on people's decision processes. Again, a single value of the threshold (a), non-decision time (t_{ND}), and drift slope (v_{slope}) are used for all prevalence conditions. Since there are only two prevalence conditions, we estimate four bias parameters (v_{50}, v_{90}, z_{50}, z_{90}). All other modeling and parameter estimation specifics are the same as with Experiment 1a.

Table 4
Estimated marginal means for Experiment 2.

RT quantile	Prevalence	Median	95% HPD	
			Lower	Upper
1	50	0.271	0.066	0.550
2	50	0.473	0.313	0.641
3	50	0.455	0.309	0.595
4	50	0.499	0.375	0.621
5	50	0.592	0.481	0.697
6	50	**0.622**	0.520	0.721
7	50	0.548	0.435	0.643
8	50	0.559	0.460	0.662
9	50	0.488	0.386	0.589
10	50	0.560	0.452	0.660
1	90	**1.000**	0.999	1.000
2	90	**0.981**	0.963	0.993
3	90	**0.971**	0.950	0.987
4	90	**0.902**	0.856	0.939
5	90	**0.836**	0.775	0.891
6	90	**0.795**	0.726	0.858
7	90	**0.665**	0.580	0.747
8	90	0.505	0.411	0.593
9	90	0.434	0.346	0.521
10	90	**0.395**	0.308	0.477

Note. Results are averaged over the levels of image type and difficulty.
Note. Results are on the response scale.
Note. Estimates in bold have 95% HPD intervals that exclude 0.50.

3.4. Modeling results

To assess the influence of prevalence on experts, we calculated the difference between the 90% prevalence parameters and the corresponding 50% prevalence parameters (e.g., $z_{90} - z_{50}$) for the start point (z) and drift (v_{prev}) parameters (Fig. 4, bottom row). As with the novice participants, results show a response expectancy bias, although the 95% CI for this difference includes 0 as shown in Table 3. Also note that the response expectancy bias for experts is smaller in magnitude than with novices (since these are measured as percentages, it is possible to directly compare their magnitude between novices and experts). Interestingly, the experts also exhibit a significant stimulus evaluation bias in the high prevalence condition (95% CI for this difference excludes 0 as shown in Table 3). This suggests that a high prevalence of blast images alters experts' evaluation of those images in addition to causing a modest response expectancy bias.

3.5. Conclusion

Similar to Experiment 1, high prevalence lead to an increase in false alarms in experts. CNN-DDM results show that prevalence effects both response expectancy and stimulus evaluation biases in experts. Whereas novices show a more pronounced response expectancy bias, experts show a more pronounced stimulus evaluation bias.

4. General discussion

Understanding why extreme prevalence rates lead to increased errors is an important step in developing ways to improve decision-making in critical real-world tasks such as baggage screening by TSA officers and the detection of abnormalities by radiologists and pathologists. While the "prevalence effect" is well established (Horowitz, 2017; Wolfe et al., 2005; Wolfe et al., 2007), its source is not well understood. One reason for this is that a common theoretical tool used to quantify its presence, Signal Detection Theory (SDT), cannot distinguish between different biases that can cause this effect. Here we demonstrated how an evidence accumulation modeling framework could be used to dissect those biases.

There are two types of biases (as conceptualized in the DDM, Leite & Ratcliff, 2011; White & Poldrack, 2014) that may arise in extreme prevalence scenarios: response expectancy and stimulus evaluation biases. The former captures biases in response preparation while the latter captures biases in the evaluation of a stimulus. We first demonstrated the inability of SDT to distinguish between these biases; they are both encoded in the criterion parameter of SDT even though they represent two fundamentally different types of processes.

We next used a new tool, the CNN-DDM (Holmes, O'Daniels, & Trueblood, 2020), which couples a Convolutional Neural Network of image perception to a Diffusion Decision Model of decision-making to assess which of these biases may be the source of prevalence effects in our task. Fitting this model to data revealed that prevalence influences both response expectancy and stimulus evaluation biases. While novices exhibit a strong response expectancy bias and weaker stimulus evaluation bias, experts show a strong stimulus evaluation bias and weaker response expectancy bias. These results are important for two reasons. First, they illustrate an important difference in how prevalence influences novices and experts. For novice participants, who receive only minimal training on cell classification, extreme prevalence leads to a strategy of responding more often for the high base-rate category regardless of the characteristics of a specific image. Experts on the other hand exhibit a prominent stimulus evaluation bias in addition to a smaller response expectancy bias, which suggests that the evaluation of cell images changes with the base-rate. The presence of a stimulus bias in experts suggests that the evaluation of a case is dependent on information from other, independent cases. This suggests that there is a relative component to information processing in experts' diagnostic decisions. The distinction between response and stimuli biases is critical as

strategies for mitigating the prevalence effect would likely depend on the source of that effect.

To date, most past research on the prevalence effect has used visual search tasks (Wolfe et al., 2005; Wolfe et al., 2007; Wolfe & Van Wert, 2010). Because search tasks generally require multiple shifts of attention before a target is found, the processing demands are very different from the image discrimination task used in this paper. It is possible that prevalence has a different influence in classification than visual search. For example, in Wolfe and Van Wert (2010), participants performed a simulated baggage search task where targets were weapons. In their task, they found that high prevalence (98% prevalence) resulted in an increase in false-alarms and a decrease in misses as compared to equal prevalence (50% prevalence). However, the prevalence manipulation had no influence on the false-alarm RTs, unlike our task. The only effect on RTs was a significant slowing of target-absent responses.

Wolfe and Van Wert (2010) suggested that the prevalence effect in visual search could be modeled using a dual-threshold model (called the "multiple-decision model") where target-present decisions are modeled using signal detection theory and search termination decisions are modeled by an accumulator model with a "quitting threshold". Wolfe and Van Wert (2010) suggested that prevalence influences both the criterion in signal detection theory as well as the "quitting threshold" of the accumulator model. For example, at high prevalence, the criterion becomes more liberal and the quitting threshold is higher. Thus, participants are more likely to identify items as targets and less likely to quit search. These changes in thresholds explain why false alarm rates are high (because of the reduced criterion) and target-absent RTs are long (because of the higher quitting threshold) in high prevalence. Note that the time to find a target is indifferent to prevalence, thus false-alarm RTs do not differ from high and equal prevalence.

Our task differs from the visual search task of Wolfe and Van Wert (2010) in several important ways. If we consider target search and target classification as two different axes of complexity, our task involves somewhat complex classification with no real search. In some cases, non-blast cells look morphologically very similar to blast cells, making the categorizing task difficult. In Wolfe and Van Wert (2010), the search process was challenging but the classification task was relatively simple (i.e. weapon versus non-weapon). The difficulty in their task arose from the number of search items (ranging from 3 to 18). It is possible the differences in RT effects (e.g., increased false alarms for fast responses in high prevalence) in our domain versus visual search are a result of the different task demands. For tasks that combine a complex visual categorization task with a search task (e.g., a baggage search task where weapons are disguised as regular travel items), the Wolfe and Van Wert (2010) model could be generalized by assuming that the identification decision is modeled by the DDM rather than SDT. This would allow the model to account for more nuanced target-present RT distributions. It would also allow researchers to model both response expectancy and stimulus evaluation biases, as in the current work.

We believe that dynamic models of the prevalence effect, such as the CNN-DDM or the "multiple-decision model" of Wolfe and Van Wert (2010), could aid researchers interested in prevalence interventions aimed at reducing errors. For example, Wolfe et al. (2007) found that in low prevalence settings, a burst of trials at 50% prevalence with feedback significantly reduced prevalence effects. Alternatively, the strategy of instructing participants to go slow or be more careful (including offering speeding tickets for very quick decisions; Solman, Wu, Cheyne, & Smilek, 2013; Wolfe et al., 2007) is ineffective in reducing the prevalence effect. This latter approach would be expected to work if a response expectancy bias is the main source of the prevalence effect, since the longer deliberation times would reduce the effect of initial bias. However, our results show that prevalence can influence both response expectancy and stimulus evaluation biases. Thus, interventions that target only one of these biases are less likely to be effective. The approach of "bursting" trials at 50% prevalence is likely more effective because it helps mitigate both types of biases. First, it would help recalibrate people's evaluation of images, thus helping to reduce a stimulus evaluation bias. Second, it could also help reset initial biases to zero by making the response strategy of selecting the higher base rate category less effective. While this likely over-simplifies the potential effects of these specific interventions, it none-the-less illustrates the need to understand the source of the prevalence effect when designing interventions.

In this paper, we used a simplified version of a real medical image interpretation task. While our task was not as complex as the real-world task, we believe our findings and methodology could have important applications to the real world. A first step in diagnosing malignant blood diseases, such as leukemia and lymphoma, is a morphologic classification of peripheral blood cells. Often, this step is performed by automated image analysis in combination with human observers. A hematology analyzer (e.g. CellaVisionTM DM96) performs the acquisition and pre-classification of cells, which are then verified or re-classified by a knowledgeable human observer (generally, a medical laboratory professional). Verification / re-classification is performed on the CellaVision's interface, where digital images of cells are presented in groups based on the machine's pre-classification. The pre-classification and grouping of cells by the analyzer have the potential to influence how observers process information and make decisions. In particular, when similar cells are grouped together by the machine, this leads to a high prevalence situation for the grouped cells. That is, if the machine groups several cells together and labels them as "blast" cells, then most of the cells in that group will be blast cells because the machine is reasonably accurate. However, the machine makes sufficient errors as to require verification / re-classification by knowledgeable humans. In this example, knowledgeable humans would need to identify any non-blast cells in the initial blast cell grouping by the machine. Thus, the machine manipulates prevalence through the grouping and presentation of images. In addition, the pre-labeling of cells by the machine could also bias human observers. Models such as the CNN-DDM can help understand the impact of automated information (e.g., grouping and pre-labeling) on human observers by disentangling latent cognitive processes, such as response expectancy and stimulus evaluation biases.

The use of artificial intelligence (AI) systems in medical image perception is not limited to classification of peripheral blood cells. In particular, pathology is on the brink of a fundamental change in how images are interpreted and diagnoses are made. Over the next decade, it is predicted that there will be a rapid rise in the use of computer-based image analysis in pathology (Granter, Beck, & Papke Jr., 2017; Sharma & Carter, 2017; Tizhoosh & Pantanowitz, 2018). This predicted rise in digital pathology is in large part due to the approval of the first whole slide imaging system for primary diagnosis in pathology by the US Food and Drug Administration in 2017 (A. J. Evans et al., 2018). As computer-based image analysis systems become more prominent in pathology, it is critical to understand how observers process automated information in order to develop user interfaces that lead to improved diagnostic accuracy as well as better quality of care. While there are many benefits to the computerization of diagnosis, not all automated and artificial intelligence systems lead to improved performance. For example, Computer-Aided Detection (CAD) systems in radiology have been shown to provide little, if any, improvements in breast cancer screening (Fenton, 2015; Fenton et al., 2011; Lehman et al., 2015; Sato et al., 2014). It is believed that the failure of CAD systems to impact performance is not related to system information, but how this information is presented to knowledgeable observers (Hupse et al., 2013; Nishikawa & Gur, 2014). Our modeling approach offers a method for examining the influence of contextual factors and automated information on how knowledgeable observers interpret images and make diagnoses.

There are some limitations to the current study. First, we only examined decisions in a single type of task, the classification of white blood cells. Future work could examine the generalizability of our findings to other types of tasks that involve categorizing complex visual stimuli as well as tasks that combine a complex visual categorization

task with a search task. Second, due to the limited pool of expert participants, we were only able to study the impact of high prevalence on experts. It is possible that low prevalence would impact experts differently. However, we suspect that we would see a similar influence of low prevalence on the stimulus evaluation bias parameter in the DDM since high prevalence of Blast images is equivalent to low prevalence of Non-Blast images.

In conclusion, we found that the CNN-DDM provides a more complete analysis of the impact of prevalence on errors than SDT. Importantly, it allows for the distinction between response expectancy biases and stimulus evaluation biases. Experimental and modeling results show that novice participants generally show a strong response expectancy bias and a weak stimulus evaluation bias under extreme prevalence. Experts, on the other hand, show a large stimulus evaluation bias and weaker response expectancy bias. This implies that prevalence can influence multiple types of biases and this influence is partially dependent on expertise.

Author contributions

All authors contributed to the study concept and design. The experimental program was coded by PO, under the supervision of JST. Testing and data collection was performed by PO and JST. Data analyses were performed by JST. Computational modeling was performed by WRH and JST. WRH and JST drafted the manuscript, and all authors provided critical revisions. All authors approved the final version of the manuscript for submission.

Acknowledgments

This work was supported by a Clinical and Translational Research Enhancement Award from the Department of Pathology, Microbiology, and Immunology, Vanderbilt University Medical Center. This work was also supported by NSF grant 1846764.

Appendix A. Supplementary materials

Supplementary analyses to this article can be found online at https://doi.org/10.1016/j.cognition.2021.104713. All of the data are available at https://osf.io/4n7sr/

References

Dutilh, G., Vandekerckhove, J., Forstmann, B. U., Keuleers, E., Brysbaert, M., & Wagenmakers, E. J. (2012). Testing theories of post-error slowing. *Attention Perception & Psychophysics, 74*(2), 454–465. https://doi.org/10.3758/s13414-011-0243-2.

Evans, A. J., Bauer, T. W., Bui, M. M., Cornish, T. C., Duncan, H., Glassy, E. F., … Myers, C. (2018). US Food and Drug Administration approval of whole slide imaging for primary diagnosis: A key milestone is reached and new questions are raised. *Archives of Pathology & Laboratory Medicine, 142*(11), 1383–1387.

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.

Evans, N. J., Holmes, W. R., & Trueblood, J. S. (2019). Response-time data provide critical constraints on dynamic models of multi-alternative, multi-attribute choice. *Psychonomic Bulletin & Review, 26*(3), 901–933.

Evans, N. J., Trueblood, J. S., & Holmes, W. R. (2019). A parameter recovery assessment of time-variant models of decision-making. *Behavior Research Methods*, 1–14.

Fenton, J. J. (2015). Is it time to stop paying for computer-aided mammography? *JAMA Internal Medicine, 175*(11), 1837–1838.

Fenton, J. J., Abraham, L., Taplin, S. H., Geller, B. M., Carney, P. A., D'Orsi, C., … Breast Cancer Surveillance, C. (2011). Effectiveness of computer-aided detection in community mammography practice. *Journal of the National Cancer Institute, 103*(15), 1152–1161.

Granter, S. R., Beck, A. H., & Papke, D. J., Jr. (2017). AlphaGo, deep learning, and the future of the human microscopist. *Archives of Pathology & Laboratory Medicine, 141*(5), 619–621. https://doi.org/10.5858/arpa.2016-0471-ED.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (Vol. 1). New York: Wiley.

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

Holmes, W. R. (2015). A practical guide to the Probability Density Approximation (PDA) with improved implementation and error characterization. *Journal of Mathematical Psychology, 68-69*, 13–24. https://doi.org/10.1016/j.jmp.2015.08.006.

Holmes, W. R., O'Daniels, P., & Trueblood, J. S. (2020). A joint deep neural network and evidence accumulation modeling approach to human decision-making with naturalistic images. *Computational Brain and Behavior, 3*(1), 1–12.

Holmes, W. R., & Trueblood, J. S. (2017). Bayesian analysis of the piecewise diffusion decision model. *Behavior Research Methods*. https://doi.org/10.3758/s13428-017-0901-y.

Holmes, W. R., Trueblood, J. S., & Heathcote, A. (2016). A new framework for modeling decisions about changing information: The Piecewise Linear Ballistic Accumulator model. *Cognitive Psychology, 85*, 1–29. https://doi.org/10.1016/j.cogpsych.2015.11.002.

Horowitz, T. S. (2017). Prevalence in visual search: From the clinic to the lab and back again. *Japanese Psychological Research, 59*(2), 65–108.

Hupse, R., Samulski, M., Lobbes, M. B., Mann, R. M., Mus, R., den Heeten, G. J., … Karssemeijer, N. (2013). Computer-aided detection of masses at mammography: Interactive decision support versus prompts. *Radiology, 266*(1), 123–129.

JASP Team. (2020). *JASP (version 0.14.1) [computer software]*.

Kruschke, J. K. (2010). Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science, 1*(5), 658–676.

Lee, M. D., Criss, A. H., Devezer, B., Donkin, C., Etz, A., Leite, F. P., … White, C. N. (2019). Robust modeling in cognitive science. *Computational Brain & Behavior, 2*(3–4), 141–153.

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

Lehman, C. D., Wellman, R. D., Buist, D. S. M., Kerlikowske, K., Tosteson, A. N. A., & Miglioretti, D. L. (2015). Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Internal Medicine, 175*(11), 1828–1837.

Leite, F. P., & Ratcliff, R. (2011). What cognitive processes drive response biases? A diffusion model analysis. *Judgment and Decision making, 6*(7), 651–687.

Lin, Y. S., Heathcote, A., & Holmes, W. R. (2019). Parallel probability density approximation. *Behavior Research Methods, 51*, 2777–2799.

Navarro, D. J., & Fuss, I. G. (2009). Fast and accurate calculations for first-passage times in Wiener diffusion models. *Journal of Mathematical Psychology, 53*(4), 222–230.

Nishikawa, R. M., & Gur, D. (2014). CADe for early detection of breast cancer—Current status and why we need to continue to explore new approaches. *Academic Radiology, 21*(10), 1320–1321.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Paper presented at the proceedings of the 3rd international workshop on distributed statistical computing*.

Ratcliff, R. (1978). Theory of memory retrieval. *Psychological Review, 85*(2), 59–108. https://doi.org/10.1037//0033-295x.85.2.59.

Ratcliff, R. (1985). Theoretical interpretations of the speed and accuracy of positive and negative responses. *Psychological Review, 92*(2), 212.

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review, 111*(2), 333–367. https://doi.org/10.1037/0033-295X.111.2.333.

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences, 20*(4), 260–281. https://doi.org/10.1016/j.tics.2016.01.007.

Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review, 106*(2), 261.

Sato, M., Kawai, M., Nishino, Y., Shibuya, D., Ohuchi, N., & Ishibashi, T. (2014). Cost-effectiveness analysis for breast cancer screening: Double reading versus single+CAD reading. *Breast Cancer, 21*(5), 532–541.

Sharma, G., & Carter, A. (2017). Artificial intelligence and the pathologist: Future frenemies? *Archives of Pathology & Laboratory Medicine, 141*(5), 622–623. https://doi.org/10.5858/arpa.2016-0593-ED.

Solman, G. J., Wu, N., Cheyne, J. A., & Smilek, D. (2013). In manually-assisted search, perception supervises rather than directs action. *Experimental Psychology, 60*(4), 243–254.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., … Rabinovich, A. (2015). Going deeper with convolutions. In *Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition*.

Tizhoosh, H. R., & Pantanowitz, L. (2018). Artificial intelligence and digital pathology: Challenges and opportunities. *Journal of Pathology Informatics, 9*.

Trueblood, J. S., Heathcote, A., Evans, N., & Holmes, W. R. (2021). Urgency, leakage, and the relative nature of information processing in decision making. *Psychological Review, 128*(1), 160–186.

Trueblood, J. S., Holmes, W. R., Seegmiller, A. C., Douds, J., Compton, M., Szentirmai, E., … Eichbaum, Q. (2018). The impact of speed and bias on the cognitive processes of experts and novices in medical image decision-making. *Cognitive Research: Principles and Implications, 3*(1), 28.

Turner, B. M., & Sederberg, P. B. (2012). Approximate Bayesian computation with differential evolution. *Journal of Mathematical Psychology, 56*(5), 375–385. https://doi.org/10.1016/j.jmp.2012.06.004.

Turner, B. M., & Sederberg, P. B. (2014). A generalized, likelihood-free method for posterior estimation. *Psychonomic Bulletin & Review, 21*(2), 227–250. https://doi.org/10.3758/s13423-013-0530-0.

White, C. N., & Poldrack, R. A. (2014). Decomposing bias in different types of simple decisions. *Journal of Experimental Psychology-Learning Memory and Cognition, 40*(2), 385–398. https://doi.org/10.1037/a0034851.

Witt, J. K., Taylor, J. E. T., Sugovic, M., & Wixted, J. T. (2015). Signal detection measures cannot distinguish perceptual biases from response biases. *Perception, 44*(3), 289–300.

Wolfe, J. M., Horowitz, T. S., & Kenner, N. M. (2005). Rare items often missed in visual searches. *Nature, 435*(7041), 439–440.

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