Investigating the Cognitive Processes Involved in Cancer Cell Image Identification

Jennifer S. Trueblood\textsuperscript{a,1,2} (jennifer.s.trueblood@vanderbilt),
William R. Holmes\textsuperscript{b,1,2} (william.holmes@vanderbilt.edu),
Adam C. Seegmiller\textsuperscript{c} (adam.seegmiller@vanderbilt.edu),
Jonathan Douds\textsuperscript{c} (jonathan.j.douds@vanderbilt.edu),
Margaret Compton\textsuperscript{c} (margaret.l.compton@vanderbilt.edu),
Megan Woodruff\textsuperscript{a} (megan.e.woodruff@vanderbilt.edu),
Wenrui Huang\textsuperscript{a} (wenrui.huang@vanderbilt.edu),
Charles Stratton\textsuperscript{c} (charles.stratton@vanderbilt.edu),
Quentin Eichbaum\textsuperscript{c,1} (quentin.eichbaum@vanderbilt.edu)

\textsuperscript{a}Department of Psychology, Vanderbilt University
\textsuperscript{b}Department of Physics and Astronomy, Vanderbilt University
\textsuperscript{c}Vanderbilt Pathology Education Research Group (VPERG), Department of Pathology, Microbiology and Immunology, Vanderbilt University Medical Center (VUMC).

\textsuperscript{1} Correspondence may be addressed to Jennifer S. Trueblood, Department of Psychology, Vanderbilt University, Nashville TN 37235, USA; Email: jennifer.s.trueblood@vanderbilt.edu or William R. Holmes, Department of Physics and Astronomy, Vanderbilt University, Nashville TN 37235, USA; Email: william.holmes@vanderbilt.edu or Quentin Eichbaum, Department of Pathology, Microbiology and Immunology, Vanderbilt University Medical Center, Nashville, TN 37232; Email: quentin.eichbaum@vanderbilt.edu

\textsuperscript{2} Authors contributed equally
Abstract

The ability to make timely and accurate decisions is critical for cancer detection in diagnostic pathology. We describe a joint experimental and computational modeling approach for determining the impact of three factors on the identification of peripheral blood blast cells (cancer indicators) versus normal white blood cells in microscopy images of clinical samples: (1) image difficulty, (2) time pressure, and (3) externally imposed bias. Our study also examines how performance on accurate identification of blast cell images is related to general object recognition ability (as measured using the Novel Object Memory Test (NOMT) (Richler et al., 2017). For this study we collected a bank of hundreds of digital images that were identified by cell type and classified by difficulty by a panel of expert hematopathologists. Participants in this study were a non-expert, student population. The Diffusion Decision Model was fit to choice and response time results of this experiment to determine how these factors influence decision processes. Results indicate that external time pressure leads to a reduction in caution in participants while external bias has a small but detectable influence on choices. Comparison of cancer image identification results with NOMT results indicates participants with stronger general object recognition ability perform better on the cell identification task, but only on easier images. This indicates that general object recognition ability carries over to this diagnostic task to a limited extent.
The ability to classify and interpret medical images is critical in the diagnosis of many diseases. Despite significant improvements in imaging assays as well as meticulous training and education, diagnostic errors still occur. In order to improve diagnostic decision-making based on medical images, it is critical to understand the cognitive processes involved in these decisions. This research borrows well-validated experimental and computational methods from perceptual decision-making and applies them to cancer cell image identification. Using a non-expert student population, we examine the impact of difficulty, time pressure, and externally imposed bias on the identification of single cell images related to cancer diagnosis. Using computational modeling techniques, we find that these manipulations have important impacts on diagnostic decisions. In addition, we find that participants with better general object recognition ability perform better at the task. In sum, these results shed light on the cognitive mechanisms that play a role in medical image perception and decision-making. In the future, this knowledge could be used to inform training and education. Moreover, this approach could be used to further investigate how experts perceive and interpret medical images.

**Keywords:** medical decision-making, diagnostic pathology, diffusion decision model, general object recognition, cancer image detection
Introduction

Accurate interpretation and classification of medical images is an important step in the diagnosis and treatment of numerous diseases. A wide range of medical disciplines (Samei & Krupinski, 2010) ranging from pathology (our focus here), to radiology, to ophthalmology rely on expert analysis of images to detect abnormalities. While the exact rate of diagnostic errors is unknown, consistent evidence suggests error rates are >10% (Goldman et al., 1983; Hoff, 2013; Kirch & Schafii, 1996; Shojania, Burton, McDonald, & Goldman, 2003; Sonderegger-Iseli, Burger, Muntwyler, & Salomon, 2000). It is thus critical that we understand how people make perceptual decisions from medical images in order to improve training and minimize the occurrence of misdiagnoses. This requires investigation of the cognitive processes underlying decision-making in this domain and how those processes are manipulated by external factors. For example, how does time pressure influence behavior? Alternatively, how do externally imposed biases influence performance? After all, in many cases, images have passed through a computer aided diagnosis process and / or analysis by medical technologists and residents before being verified by a senior clinical expert.

Decisions based on medical images have a number of parallels with perceptual decision-making, where people make choices based on sensory information. The investigation of perceptual decision-making has a rich history in psychology, cognitive science, and neuroscience (Gold & Shadlen, 2007; Heekeren, Marrett, & Ungerleider, 2008; Summerfield & de Lange, 2014; Summerfield & Egner, 2009). However, perceptual decision-making of medical images in clinical settings has received less attention. This is in part due to several logistical challenges not typically present in lab
settings (Wolfe, 2016). First, the collection of an annotated bank of images requires significant time investment from experts to classify those images. Second, time and access become issues if you are looking to experts to participate in experimental studies: there are usually few local experts in a particular diagnostic specialty (low sample size) and they are usually very busy. Third, manipulating cognitive variables or aspects of the experiment can be challenging. Thus it is important to devise strategies for investigating this domain in both clinical and non-clinical settings, and to find ways to connect the larger body of prior work on lab based perceptual decision-making to the clinical context.

Here, we present a study investigating the cognitive processes underlying cancer image detection in diagnostic pathology. More specifically, we investigate how various external factors influence the ability of non-expert undergraduate students to distinguish between normal (standard white blood cells such as lymphocytes, monocytes, or neutrophils) and abnormal cancer cells (“blast” cells, associated with acute leukemia) in clinical images. Toward this end, we take a joint experimental and modeling approach utilizing experimental paradigms and modeling methods previously developed in the course of basic research on perceptual decision-making (Ratcliff & Smith, 2004; Schouten & Bekker, 1967; Wickelgren, 1977).

To investigate this process experimentally, we passively collected a large bank of digital images of both blast and non-blast white blood cells drawn from patients at the Vanderbilt University Medical Center (all images were obtained as part of routine clinical care). A panel of expert pathologists classified each of these images, providing a fully annotated data set consisting of hundreds of images of varying type and level of difficulty. Using this image bank, we developed a speeded decision-making experiment
to investigate how difficulty, time pressure, and externally imposed bias influence individuals’ behavior.

In addition to testing participants’ ability to discriminate between and classify images of blast and non-blast cells, we also investigate how participants’ general object recognition ability impacts their performance on this task. Toward this end, we employ a second task, the Novel Object Memory Test (NOMT), to assess each participant’s general ability to learn and recognize objects that they have no prior experience with (Richler, Wilmer, & Gauthier, 2017). We use this to probe to what extent general object recognition, which has been studied in much more detail in lab settings (Gauthier et al., 2014; Hildebrandt, Wilhelm, Herzmann, & Sommer, 2013; McGugin, Gatenby, Gore, & Gauthier, 2012; McGugin, Richler, Herzmann, Speegle, & Gauthier, 2012), correlate with or impact participants’ efficacy on the blast cell identification task.

To gain further insight into the cognitive processes underlying decisions on this task, we utilize computational modeling linked with results of this experiment. One of the benefits of quantitative modeling, and the reason we use it here, is that it provides a way to quantify latent cognitive processes and statistically separate the different components of the decision process (caution, bias, and rate of information uptake) that are not accessible through traditional statistical methods alone. For this, we utilize a version of the classic Diffusion Decision Model (DDM) (Ratcliff, 1978; Ratcliff & McKoon, 2008; Ratcliff, Smith, Brown, & McKoon, 2016), which has been shown to account for detailed patterns of behavior across a wide range of speeded decision-making paradigms (Ratcliff, Love, Thompson, & Opfer, 2012; Ratcliff, Thapar, & McKoon, 2001, 2004, 2010), to model the choice and response time behavior of participants on this task and extract these
underlying cognitive parameters.

**Experimental Methods**

*Participants.* 37 undergraduate students at Vanderbilt University participated in exchange for course credit. We set a target sample size of about 35 participants. This sample size was selected on the basis of previous experiments conducted by the authors that used similar modeling methods (Holmes, Trueblood, & Heathcote, 2016).

*Materials.* To create the stimuli, we collected a bank of 840 digital images of Wright-stained white blood cells taken from anonymized patient peripheral blood smears at Vanderbilt University Medical Center. The images were taken by a CellaVision DM96 automated digital cell morphology instrument (CellaVision AB, Lund, Sweden). This instrument and accompanying software identifies and images single white blood cells, and classifies them into one of 17 cell types based on morphologic characteristics. The classification of each cell is confirmed by a trained medical technologist.

A ratings panel of three hematopathology faculty from the Department of Pathology at Vanderbilt University Medical Center was used to identify and rate each image. The raters first identified each image as a blast or a non-blast cell. Following this identification, they were asked to provide a difficulty rating for each image on a 1-5 scale. If the raters identified the image as a blast cell, they were asked to rate how similar the image was to a classic blast cell (with a rating of 1 being ‘not similar’ and a rating of 5 being ‘very similar’). Raters were told that a classic blast cell image is one that might be used in a textbook. If raters identified the image as a non-blast cell, they were asked to
rate how morphologically similar the cell is to a blast cell (with a rating of 1 being ‘not similar’ and a rating of 5 being ‘very similar’).

Out of the original set of 840 images, the three expert raters agreed on the cell type (i.e., blast or non-blast) for 633 images. From this set, we grouped the images into four types based on the difficulty ratings. The average rating for the blast images was 4.40 (SD= 0.46) and the average rating for the non-blast images was 1.68 (SD = 0.55). Blast images that received an average rating of 4.5 or greater were categorized as easy blasts (151 images, Fig 1a). Blasts images that received an average rating of 4 or less were categorized as hard blasts (98 images, Fig 1b). Images with an average rating between 4 and 4.5 were not included. Non-blast images that received an average rating of 1.5 or less were categorized as easy non-blasts (129 images, Fig 1c). Non-blasts images that received an average rating of 2 or greater were categorized as hard non-blasts (108 images, Fig 1d). Images with an average rating between 1.5 and 2 were not included. For the experiment, we selected 75 images from each category for a total of 300 unique images. Note that we have more than 300 trials in the experiment, so some images are repeated. However, no images were repeated until all of the images from a category had been shown. Thus, we doubt participants were able to recognize that images were shown more than once.

**Procedure.** In the main task, participants first completed a training stage to familiarize themselves with blast cells. There were four blocks of training trials. Each block started with participants studying five images of blast cells one at a time. After studying these five images, participants then completed 15 trials where they were presented three images (1 blast image and 2 non-blast images) and asked to choose the
image they thought was the blast cell. The four training blocks had the following structure of blast and non-blast images: block 1 was easy blast versus easy non-blast, block 2 was easy blast versus hard non-blast, block 3 was hard blast versus easy non-blast, and block 4 was hard blast versus hard non-blast.

After completing the four training blocks, participants completed a practice block of 60 trials to familiarize themselves with the main task. Each trial started with a fixation cross displayed for 250 ms. After fixation, participants were shown a single image and had to identify it as a blast or non-blast cell. Participants received trial-by-trial feedback about their choices in this block.

The main task consisted of six blocks with 100 trials in each block. The main task was the same as the practice block where participants were asked to identify single images. However, participants did not receive trial-by-trial feedback about their choices. They received feedback about their performance at the end of each block. The 100 trials in each block were composed of equal numbers of easy blast images, hard blast images, easy non-blast images, and hard non-blast images, fully randomized.

There were three manipulations across blocks: accuracy, speed, and bias. In the accuracy blocks, participants were instructed to respond as accurately as possible and were given 5 seconds to respond. In the speed block, participants were instructed to respond quickly and were given 1 second to respond. If they responded after the deadline, they received the message “Too Slow!” In the bias blocks, participants were shown a probabilistic cue on half of the trials. The cue was a red dot that was shown after fixation for 500 ms. The cue identified the upcoming image as most likely being a blast cell. The cue was valid 65% of the time. Participants were instructed about the validity of the cue
at the start of the block. The order of the first three blocks was randomized, but with the constraint that there was one block for each type of manipulation (i.e., accuracy, speed, and bias). The order of the last three blocks was identical to the order of the first three blocks.

After completing the main task, participants completed a version of the Novel Object Memory Test (NOMT) (Richler et al., 2017). The NOMT is modeled after the Cambridge Face Memory Test (Duchaine & Nakayama, 2006) and provides a measure of general object recognition ability. In our experiment, we used two categories of novel objects (Ziggerins shown in Fig 1e and Greebles shown in Fig 1f). For each category, participants started with a learning phase where a target object was shown in three views followed by three test trials where the target was shown alongside two distractor objects. Participants received trial-by-trial feedback during these trials. This learning procedure was repeated for 6 target objects (the six targets for each category are shown in Fig 1e and 1f). Each target object was slightly different from the other targets in the same category. Following the learning phase, participants completed 54 test phase trials where they had to select which of three objects was any one of the six studied targets.

**Modeling Methods**

To gain insight into participants decision process beyond what is possible with statistical analysis of behavioral results alone, we use the canonical Diffusion Decision Model (DDM) of speeded decision making. The DDM posits that over the course of a decision, evidence stochastically accumulates over time until a confidence threshold is
reached and a decision is initiated. This model has three core elements that makes it valuable in assessing participants behavior on the blast identification task. 1) How fast people accumulate evidence over time is encoded in an *accumulation rate* parameter \( (d) \). A high / low rate indicates better / worse performance on the task. 2) The level of confidence a person requires to make a decision (i.e. level of caution) is encoded in a *threshold* parameter \( (a) \). 3) Finally, any initial preference for responding one way or the other prior to seeing the stimulus is described by a *bias* parameter \( (z) \). These are the three critical characteristics / parameters in the model that we will rely on to make inferences.

See Figure 2 for a schematic description of the DDM.

The full version of the DDM that we use here is comprised of 9 (or 10) free parameters: accumulation rates for easy and hard blast images \( (d_{BE}, d_{BH}) \), accumulation rate for easy and hard non-blast images \( (d_{NE}, d_{NBH}) \), trial to trial variability in those accumulation rates \( (s_d) \), start point \( (z\), which determines the initial bias), trial to trial variability in the start point \( (s_z) \), evidence threshold \( (a) \), encoding and response time \( (t_{ND}) \). There is also a parameter encoding within trial stochasticity \( (s) \). However as is common, we fix this parameter to a value \( s=0.1 \) to avoid parameter degeneracy in the model (one parameter must be fixed). For the cueing instruction data, we introduce an additional parameter to denote the bias on trials where the cue is actually shown \( (z_{cue}) \). This will allow us to determine if the cue has any discernible effect on initial bias. Given that the speed, accuracy, and cueing instruction conditions all have the potential to influence people’s behavior in different ways, we fit each instruction condition separately and do not assume up front that any model parameters are the same across experimental conditions.
We use a hierarchical Bayesian algorithm to fit DDM to all participants’ data simultaneously, providing an account of the choices made and the full distribution of response times at both the individual and population levels. Given the high level of correlation between model parameters in this model, we utilize Differential Evolution Markov Chain Monte Carlo (DEMCMC) (Turner & Sederberg, 2012) to carry out this Bayesian estimation. Since the DDM does not have an analytically tractable closed form likelihood function, we utilize a recently developed approximation, the Probability Density Approximation (PDA) method (Holmes, 2015; Holmes & Trueblood, 2017; Turner & Sederberg, 2014), to approximate the likelihood of each parameter set sampled.

**Behavioral Results and Discussion**

We first examined average accuracy on the 60 practice trials preceding the main blast identification task to see how well participants learned to identify the images. The proportion of trials answered correctly in the practice block was 0.73 (SD = 0.09). We removed three participants with accuracy less than two standard deviations below the average because these participants were likely not engaged in the task.

We use Bayesian statistics for all analyses in this paper. All tests were implemented using the open source software package JASP (Team, 2016). For each test, we report the Bayes factor (BF), which is the ratio quantifying the evidence in the data favoring one hypothesis relative to another. While BFs are directly interpretable, labels for the strength of Bayes factors have been proposed. In particular, BF greater (less) than
1, 3 (1/3), 10 (1/10), 30 (1/30), and 100 (1/100) are considered ‘Anecdotal’, ‘Moderate’, ‘Strong’, ‘Very Strong’, and ‘Extreme’, evidence, respectively (Kass & Raftery, 1995).

We compared the proportion of correct responses in the accuracy and speed blocks for the four image types using a Bayesian Repeated Measures ANOVA. For this analysis, we compared five models: (1) a null model, (2) a model with only a main effect of block, (3) a model with only a main effect of image type, (4) a model with both a main effect of block and image type, and (5) a model with both main effects and their interaction. We calculated the BF's comparing each model with all other possible models (BF\text{Model}) as well as with the null model (BF_{10}). The model including both main effects (and not their interaction) was preferred to all other models (BF_{Model} = 11.39) as well as to the null model (BF_{10} > 1,000). We also calculated the BF's for the inclusion of specific variables (BF_{inclusion}). The BF for inclusion of image type was large, BF_{inclusion} > 1,000, but the BF for inclusion of block was much smaller, BF_{inclusion} = 2.39. This suggests that participants’ ability to correctly identify images was more strongly influenced by image type than speed versus accuracy instructions. Figure 3a shows the proportion of correct choices for speed and accuracy blocks for the four image types. As can be seen in the figure, participants made more correct identifications for the easy images as compared to the hard images. Performance was the worst for the hard non-blast images.

We used a similar analysis to compare the mean response times in the accuracy and speed blocks for the four image types. The model including both main effects (block and image type) was preferred to all other models (BF_{Model} = 2.83) as well as to the null model (BF_{10} > 1,000). The BF for inclusion of block was large, BF_{inclusion} > 1,000, but the BF for inclusion of image type was much smaller, BF_{inclusion} = 1.41. This suggests that
accuracy versus speed instructions had the largest effect on mean response times. Figure 3c shows the mean response times for speed and accuracy blocks for the four image types. As illustrated in the figure, participants responded more quickly in the speed blocks than the accuracy blocks.

Next we examined the proportion of correct responses in the bias condition when the cue was present versus absent for the four image types using a Bayesian repeated Measures ANOVA. Similar to the above analyses, we compared five models: (1) a null model, (2) a model with only a main effect of cue presence, (3) a model with only a main effect of image type, (4) a model with both a main effect of cue presence and image type, and (5) a model with both main effects and their interaction. The model including only the main effect of image type was preferred to all other models (BF_{Model} = 24.00) as well as to the null model (BF_{10} > 1,000). The BF for inclusion of image type was large, BF_{inclusion} > 1,000. The BF for inclusion of cue presence was 0.11, suggesting that the presence / absence of the cue had no impact on choice behavior. Figure 3b shows the proportion of correct choices for the bias blocks when the cue was present versus absent for the four image types. Similar to the accuracy and speed blocks, participants made more correct identifications for the easy images as compared to the hard images. Performance was the worst for the hard non-blast images.

We then compared the mean response times in the bias condition when the cue was present versus absent for the four image types. The model including only the main effect of image type was preferred to all other models (BF_{Model} = 4.15) as well as to the null model (BF_{10} > 1,000). The BF for inclusion of image type was large, BF_{inclusion} > 1,000. The BF for inclusion of cue presence was 0.64, suggesting that the presence /
absence of the cue had no impact on mean response time. Figure 3d shows the mean response times for the bias blocks when the cue was present versus absent for the four image types. As can be seen in the figure, participants responded more quickly for the easy non-blast images as compared to the other image types.

Next, we compared identification performance on the main task with performance on the NOMT. The average proportion of correct responses on the NOMT was 0.71 (SD = 0.10). We correlated NOMT performance with performance on the identification task for easy and hard trials separately. As shown in Figure 4a, we found a positive correlation between NOMT performance and performance on the easy trials of the main task (r = 0.39, BF10 = 2.57). As illustrated in Figure 4b, there was no correlation between NOMT performance and performance on the hard trials (r = 0.12, BF10 = 0.26).

**Modeling Results and Discussion**

We fit the DDM to each of the speed, accuracy, and cueing instruction conditions separately. It is in principle possible to fit the totality of the data at once as is often done. Typically this is accomplished by fixing certain parameters (accumulation rates for example) to be the same across instruction conditions while others (threshold for example) are condition dependent. This however restricts up front the properties of the model that can vary between conditions. By fitting the three conditions separately, we allow maximal model flexibility, so that the data can determine what is the same or different across conditions.
To determine if the model was able to capture the data (which is necessary for it to be useful), we 1) extracted the mean parameters for fits to each of the conditions, 2) calculated the predicted choice proportion and mean response time (RT) for each condition, and 3) compared those predictions to choice proportions and mean RT’s from data. Results (Figure 5) show that the model provides a good accounting of most aspects of the observed data. In each of these figure panels, the diagonal line represents the line of perfect agreement (prediction = observation), with results lying close to this line in most cases. There are two exceptions. First, there is more spread in the fits for the cuing condition. Note however that each of the cueing experimental conditions has $\frac{1}{2}$ as many observations as the speed and accuracy conditions (due to the presence of cue / no cue trials). Thus increased noise in the data would be expected. Second, the model does not provide a good accounting of the long response times, which is a common issue with DDM and other similar models since it is a model of speeded decision making. Overall, the model fits the observed data well and thus we will further analyze the model results.

We next look at the best-fit values for the three key model parameters linked to behavioral characteristics of interest here (accumulation rate, threshold, bias). Figure 6 shows posterior distributions (best fit parameter distributions) for these parameters in each of the three instruction conditions. Analysis of the drift rate estimates shows a consistent pattern across conditions. As expected, evidence accumulation rates are lower for more difficult images, regardless of instruction condition and difficulty. There is however a significant difference between participants’ ability to perceive the characteristics of blast and non-blast images respectively, as evidenced by the fact that $d_{NB} \neq d_B$. It appears that the characteristics of hard non-blast images are the most difficult
to discern while the characteristics of easy non-blast are the simplest. This is consistent with behavior observations (Figures 3 a,c) where participants made more errors on hard non-blast images. Further, while participants had similar numbers of correct and error responses on easy blast and non-blast images, they responded more quickly in the easy non-blast case (in the accuracy condition only).

Results additionally show that there is no detectable bias in the speed or accuracy conditions. That is, participants had no implicit preference for identifying cells as either blast or non-blast. The introduction of a cue indicating a higher likelihood that the subsequent image is a blast does appear to introduce some bias. While this bias is relatively small (~5% of the threshold), it did appear to reduce the number of errors in cued easy, blast trials (Figure 3b). Comparison of the threshold parameters between the speed and accuracy conditions suggests that the speed instruction dominantly influences the threshold parameter. Thus, under speed instructions, it appears that participants become less cautious rather than more perceptive or focused (which would show up in rates).

Next, we compared participants’ performance on the NOMT with measures of speed, accuracy, and bias derived from this modeling. To do so, we used Bayesian linear regression to predict NOMT performance using the best-fit parameters from the DDM for each individual. We carried out the linear regression analyses separately for accuracy, speed, and bias blocks since the model was fit separately to these conditions. For the accuracy and speed conditions, there were 7 predictors ($t_{NB}, d_{BE}, d_{BH}, d_{NBE}, d_{NBH}, a, \text{bias (a-z/2)}$). For the bias condition, there were 8 predictors since there were two different biases in the model (one for cued trials and one for uncued trials). We examined all
possible combinations of predictors ($2^7 = 128$ models were fit for the speed and accuracy conditions and $2^8 = 256$ models were fit for the bias condition). For the accuracy condition, the preferred model was the one with only $d_{NBE}$ and bias ($BF_{Model} = 22.38$ and $BF_{10} = 225.44$). For the speed condition, the preferred model was the one with only $d_{NBE}$ ($BF_{Model} = 10.86$ and $BF_{10} = 188.07$). For the bias condition, the preferred model was one with all drift rates ($d_{BE}$, $d_{BH}$, $d_{NBE}$, $d_{NBH}$) and both bias parameters ($BF_{Model} = 27.09$ and $BF_{10} = 346.26$). For each condition, we also calculated the $BF_{inclusion}$ of each predictor as shown in Table 1. As can be seen in the table, the $BF_{inclusion}$ for the $d_{NBE}$ parameter was large for all three conditions. In particular, participants with larger $d_{NBE}$ parameter values had better performance on the NOMT. For the accuracy and speed conditions, the $BF_{inclusion}$ for all other parameters was small (less than 3, which is typically considered ‘anecdotal’). For the bias condition, the $BF_{inclusion}$ was small for all parameters except the bias parameter for uncued trials. In particular, better NOMT performance was associated with smaller bias parameter values.

| Effects | Accuracy | Speed | Bias |
|---------|----------|-------|------|
| $t_{ND}$ | 0.425 | 0.450 | 0.559 |
| $d_{BE}$ | 0.403 | 0.414 | 1.429 |
| $d_{BH}$ | 0.425 | 0.383 | 1.207 |
| $d_{NBE}$ | 33.227 | 74.805 | 11.440 |
| $d_{NBH}$ | 0.453 | 0.897 | 1.515 |
| thres | 0.387 | 0.451 | 0.466 |
| bias | 2.519 | 1.167 | - |
| bias cue | - | - | 1.700 |
| bias no cue | - | - | 21.088 |
These results show that the primary cognitive DDM parameter that correlates with NOMT performance is the evidence accumulation rates on easy non-blast images. This suggests that a persons’ general object recognition ability transfers (to some extent) to perception of easy images, but not hard images. We posit a possible explanation for this limited correlation. There are two basic levels of information contained within these images associated with 1) gross morphological characteristics (which are more readily learned) and 2) more detailed characteristics (which require more practice and expertise). Easy blast and non-blast images could potentially be distinguished primarily based on gross morphology while discrimination of hard blast and non-blast images requires consideration of more detailed characteristics. In this case, superior general object recognition capabilities in this untrained population appears be associated with improved ability to recognize those gross characteristics, but does not provide any significant advantage where consideration of fine details (and hence expertise) are required.

**Conclusions**

In this study, we took a joint experimental and modeling approach to investigate the cognitive processes involved in cancer cell image detection in diagnostic pathology. While this study was performed with non-expert undergraduate students, we used clinical images curated by a panel of expert pathologists. Despite the lack of experience with medical image analysis, this pool of participants was able to learn the characteristics of blast (abnormal) cells and to discriminate between them and normal cells with a reasonable level of accuracy.
Using standard paradigms in perceptual decision-making (e.g., speed-accuracy tradeoff), we found that time pressure led to a reduction in caution in participants, with a commensurate increase in errors, as is common in other domains (Bogacz, Wagenmakers, Forstmann, & Nieuwenhuis, 2010; Starns & Ratcliff, 2010; Wickelgren, 1977). Additionally, we found that participants had no implicit bias to classify images as either blast or non-blast. Even an external cue, indicating an increased likelihood of the subsequent image being a blast cell, only introduced a slight bias. Finally, we analyzed participants’ ability to discriminate blast and non-blast images of different difficulty by showing images that were deemed (by expert pathologists) to be more or less difficult to identify. Unsurprisingly, participants had more trouble with difficult images. Further analysis of both behavioral and modeling results however revealed that participants had different capacities to discriminate blast and non-blast cells, as indicated by a robust ordering of evidence accumulation rates across instruction conditions. This is possibly due to the fact that there is one blast cell type while there are multiple non-blast cell types (all other white blood cell types).

In addition to considering performance on this task, we investigated how general object recognition ability impacts error rates using the NOMT. Results show that there is a positive correlation between performances on the two tasks. That correlation is however limited to performance on easy trials and the only cognitive model parameter to correlate with NOMT performance (across all three instructional manipulations) is the rate of evidence accumulation on easy non-blast trials. We hypothesize this limited correlation is the result of participants with good object recognition ability being able to better learn
and recognize gross morphological characteristics of these cell types, but not more detailed characteristics, which require further training and expertise.

In conclusion, we find that differences in the perception of different cell types (which is of course task specific), caution, and general ability to learn and recognize objects all influence performance, though to differing extents. Further, and somewhat surprisingly, this participant population exhibits essentially no bias in their decisions, and probabilistic cues intended to influence that bias appear to have little effect on it. While experimentation alone can be used to quantify and describe errors, the joint experimental and modeling approach used here provides a way to focus more closely on the cognitive factors that influence decisions. Hopefully this study and future ones will help shed light on the role of cognitive processes in medical image perception and decision-making.
Declarations

Ethics approval and consent to participate
This study and all of its materials and consent documents were approved prior to the initiation of data collection by the Vanderbilt University Institutional Review Board (IRB # 161767).

Consent for publication
Not applicable

Availability of data and material
All de-identified data are available upon requests made to the corresponding author, jennifer.s.trueblood@vanderbilt.edu

Competing interest
The authors declare that they have no competing interests.

Funding
This work was supported by a Clinical and Translational Research Enhancement Award from the Department of Pathology, Microbiology, and Immunology, Vanderbilt University Medical Center. JST and WRH were supported by National Science Foundation grant SES-1556415.
Authors’ contributions

All authors contributed to the study concept and design. The experimental program for the blast identification task was coded by MW and the experimental program for the NOMT was coded by WH, both under the supervision of JST and WRH. Testing and data collection was performed by MW, WH, 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.

Acknowledgements

The authors would like to thank Isabel Gauthier for advice on using the NOMT.
Bogacz, R., Wagenmakers, E. J., Forstmann, B. U., & Nieuwenhuis, S. (2010). The neural basis of the speed-accuracy tradeoff. *Trends Neurosci, 33*(1), 10-16. doi:10.1016/j.tins.2009.09.002

Duchaine, B., & Nakayama, K. (2006). The Cambridge Face Memory Test: results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia, 44*(4), 576-585. doi:10.1016/j.neuropsychologia.2005.07.001

Gauthier, I., McGugin, R. W., Richler, J. J., Herzmann, G., Speegle, M., & Van Gulick, A. E. (2014). Experience moderates overlap between object and face recognition, suggesting a common ability. *Journal of Vision, 14*(8). doi:Artn 7 10.1167/14.8.7

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annu Rev Neurosci, 30*, 535-574. doi:10.1146/annurev.neuro.29.051605.113038

Goldman, L., Sayson, R., Robbins, S., Cohn, L. H., Bettmann, M., & Weisberg, M. (1983). The value of the autopsy in three medical eras. *N Engl J Med, 308*(17), 1000-1005. doi:10.1056/NEJM198304283081704

Heekeren, H. R., Marrett, S., & Ungerleider, L. G. (2008). The neural systems that mediate human perceptual decision making. *Nature Reviews Neuroscience, 9*(6), 467-479. doi:10.1038/nrn2374

References
Hildebrandt, A., Wilhelm, O., Herzmann, G., & Sommer, W. (2013). Face and Object Cognition Across Adult Age. *Psychology and Aging, 28*(1), 243-248. doi:10.1037/a0031490

Hoff, S. R. (2013). Breast Cancer: Missed Interval and Screening-detected Cancer at Full-Field Digital Mammography and Screen-Film Mammography-Results from a Retrospective Review (vol 264, pg 378, 2012). *Radiology, 266*(1), 367-367. doi:10.1148/radiol.12124051

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. doi:10.1016/j.jmp.2015.08.006

Holmes, W. R., & Trueblood, J. S. (2017). Bayesian analysis of the piecewise diffusion decision model. *Behav Res Methods*. doi: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. doi:10.1016/j.cogpsych.2015.11.002

JASP Team (2016). JASP (Version 0.7.5.5)[Computer software]

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association, 90*(430), 773-795. doi:10.1080/01621459.1995.10476572

Kirch, W., & Schafii, C. (1996). Misdiagnosis at a university hospital in 4 medical eras - Report on 400 cases. *Medicine, 75*(1), 29-40. doi:10.1097/00005792-199601000-00004
McGugin, R. W., Gatenby, J. C., Gore, J. C., & Gauthier, I. (2012). High-resolution imaging of expertise reveals reliable object selectivity in the fusiform face area related to perceptual performance. *Proc Natl Acad Sci U S A, 109*(42), 17063-17068. doi:10.1073/pnas.1116333109

McGugin, R. W., Richler, J. J., Herzmann, G., Speegle, M., & Gauthier, I. (2012). The Vanderbilt Expertise Test reveals domain-general and domain-specific sex effects in object recognition. *Vision Research, 69*, 10-22. doi:10.1016/j.visres.2012.07.014

Ratcliff, R. (1978). Theory of Memory Retrieval. *Psychological Review, 85*(2), 59-108. doi:10.1037//0033-295x.85.2.59

Ratcliff, R., Love, J., Thompson, C. A., & Opfer, J. E. (2012). Children are not like older adults: a diffusion model analysis of developmental changes in speeded responses. *Child Dev, 83*(1), 367-381. doi:10.1111/j.1467-8624.2011.01683.x

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation, 20*(4), 873-922. doi:10.1162/Neco.2008.12-06-420

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review, 111*(2), 333-367. doi: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 Cogn Sci, 20*(4), 260-281. doi:10.1016/j.tics.2016.01.007
Ratcliff, R., Thapar, A., & McKoon, G. (2001). The effects of aging on reaction time in a signal detection task. *Psychology and Aging, 16*(2), 323-341.

Ratcliff, R., Thapar, A., & McKoon, G. (2004). A diffusion model analysis of the effects of aging on recognition memory. *Journal of Memory and Language, 50*(4), 408-424. doi:10.1016/j.jml.2003.11.002

Ratcliff, R., Thapar, A., & McKoon, G. (2010). Individual differences, aging, and IQ in two-choice tasks. *Cogn Psychol, 60*(3), 127-157. doi:10.1016/j.cogpsych.2009.09.001

Richler, J. J., Wilmer, J. B., & Gauthier, I. (2017). General object recognition is specific: Evidence from novel and familiar objects. *Cognition, 166*, 42-55. doi:10.1016/j.cognition.2017.05.019

Samei, S., & Krupinski, E. (2010). *The Handbook of Medical Image Perception and Techniques* (1 ed.): Cambridge University Press.

Schouten, J. F., & Bekker, J. A. (1967). Reaction time and accuracy. *Acta Psychol (Amst), 27*, 143-153.

Shojania, K. G., Burton, E. C., McDonald, K. M., & Goldman, L. (2003). Changes in rates of autopsy-detected diagnostic errors over time: a systematic review. *JAMA, 289*(21), 2849-2856. doi:10.1001/jama.289.21.2849

Sonderegger-Iseli, K., Burger, S., Muntwyler, J., & Salomon, F. (2000). Diagnostic errors in three medical eras: a necropsy study. *Lancet, 355*(9220), 2027-2031.

Starns, J. J., & Ratcliff, R. (2010). The effects of aging on the speed-accuracy compromise: Boundary optimality in the diffusion model. *Psychology and Aging, 25*(2), 377-390. doi:10.1037/a0018022
Summerfield, C., & de Lange, F. P. (2014). Expectation in perceptual decision making: neural and computational mechanisms. *Nature Reviews Neuroscience, 15*(11), 745-756. doi:10.1038/nrn3838

Summerfield, C., & Egner, T. (2009). Expectation (and attention) in visual cognition. *Trends Cogn Sci, 13*(9), 403-409. doi:10.1016/j.tics.2009.06.003

Turner, B. M., & Sederberg, P. B. (2012). Approximate Bayesian computation with differential evolution. *Journal of Mathematical Psychology, 56*(5), 375-385. doi: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. doi:10.3758/s13423-013-0530-0

Wickelgren, W. A. (1977). Speed-Accuracy Tradeoff and Information-Processing Dynamics. *Acta Psychol (Amst), 41*(1), 67-85. doi:10.1016/0001-6918(77)90012-9

Wolfe, J. M. (2016). Use-inspired basic research in medical image perception. *Cogn Res Princ Implic, 1*(1), 17. doi:10.1186/s41235-016-0019-2
Figures Captions

Figure 1: Panels a-d) Sample images of blast and non-blast cells that were classified as easy and difficult. Panel a is an easy blast image, panel b is a hard blast image, panel c is an easy non-blast image, and panel d is a hard non-blast image. Panels e,f) Two sets of sample images from the Novel Object Memory Task (NOMT). Panel e shows the six Ziggerin targets and panel f shows the six Greeble targets.

Figure 2: Diffusion Decision Model Schematic. Evidence accumulates over time based on the stimulus present. Here the top / bottom boundaries indicate the levels of evidence needed to respond blast / non-blast respectively.

Figure 3: Accuracy and Speed results from the blast identification task. Panels a, c) Proportion of correct responses (a) and mean response times (c) in the speed and accuracy instruction conditions for the four categories of images. Panels b, d) Proportion of correct response (b) and mean response time (d) in the cued and uncued conditions in the cueing instruction conditions. Error bars are the standard error of the mean.

Figure 4: Correlation between performance on blast identification and NOMT tasks. Horizontal axes indicate response accuracy on the NOMT and vertical axes response accuracy on the identification task. Lines represent the best fit regression lines. Correlations computed separately for the easy (a) and hard (b) image classes. Within each
image class, the proportion of correct responses is calculated over all instruction conditions.

Figure 5: Quality of Model Fit. Comparison of model predictions (vertical axes) and observed (horizontal axes) for response proportions (a-c) and mean response times (d-f) in the three instruction conditions. The solid diagonal line indicates perfect agreement where predictions and observations exactly coincide.

Figure 6: Model Fit Results. Posterior distributions for the threshold, bias, and evidence accumulation rate parameters (hyper mean parameters) for the three instruction conditions. Threshold and Bias estimates are in (a-c), with (c) presenting bias estimates for the cued (red) and uncued (black) conditions separately. Evidence accumulation rates are shown in (d-f). Note that in the model, drift rates for blast and non-blast images are positive and negative respectively. Here we have presented $d_{BH}$, $d_{BE}$, $-d_{NBH}$, $-d_{NBE}$ for ease of comparison.
Figure 1
Figure 2

Diffusion Decision Model

Blast Response

Threshold

Non-Blast Response

Evidence

Bias

Time
Figure 3

a) Proportion Correct

b) Proportion Correct

c) Mean RT (sec)

d) Mean RT (sec)
Figure 4

a) Proportion Correct

Easy

r = 0.39

b) Proportion Correct

Hard

r = 0.12
Figure 5

(a) Accuracy
(b) Speed
(c) Cueing

d) Computed Choice Prob vs Data Choice Prob

Accuracy
- Hard
- Easy

Speed

Cueing
- Cue
- No Cue
Figure 6

a) **Accuracy**

b) **Speed**

c) **Cueing**

d) **Accumulation Rates**

- **Threshold Bias**
- **Cued Bias**