Cognitive Psychology 103 (2018) 1–22

Contents lists available at ScienceDirect

Cognitive Psychology

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

Modeling 2-alternative forced-choice tasks: Accounting for both magnitude and difference effects

Roger Ratcliff[a,*], Chelsea Voskuilen[a], Andrei Teodorescu[b]

[a] The Ohio State University, United States
[b] University of Haifa, Israel

ARTICLE INFO

Keywords:
Response time models
Diffusion model
LCA
2AFC task

ABSTRACT

We present a model-based analysis of two-alternative forced-choice tasks in which two stimuli are presented side by side and subjects must make a comparative judgment (e.g., which stimulus is brighter). Stimuli can vary on two dimensions, the difference in strength of the two stimuli and the magnitude of each stimulus. Differences between the two stimuli produce typical RT and accuracy effects (i.e., subjects respond more quickly and more accurately when there is a larger difference between the two). However, the overall magnitude of the pair of stimuli also affects RT and accuracy. In the more common two-choice task, a single stimulus is presented and the stimulus varies on only one dimension. In this two-stimulus task, if the standard diffusion decision model is fit to the data with only drift rate (evidence accumulation rate) differing among conditions, the model cannot fit the data. However, if either of one of two variability parameters is allowed to change with stimulus magnitude, the model can fit the data. This results in two models that are extremely constrained with about one tenth of the number of parameters than there are data points while at the same time the models account for accuracy and correct and error RT distributions. While both of these versions of the diffusion model can account for the observed data, the model that allows across-trial variability in drift to vary might be preferred for theoretical reasons. The diffusion model fits are compared to the leaky competing accumulator model which did not perform as well.

1. Introduction

The majority of the decision-making literature has focused on tasks involving categorization of a single stimulus into one of two response options. As such, most of the modeling work in this domain has focused on this type of task and the behavioral findings associated with it. In this article we examine performance in five perceptual 2-alternative forced-choice (2AFC) tasks in which two stimuli are presented and the task is to make a comparative judgment (e.g., which of the two patches contains more white pixels). In these 2AFC tasks, the representation of evidence used to make a decision can differ on two dimensions, namely the difference between the stimuli and the magnitudes of the two stimuli. This added degree of freedom in the manipulation of stimulus values can result in behavioral effects that are not observed in tasks with only a single stimulus (Teodorescu & Usher, 2013). We briefly describe the differences between 2AFC tasks and standard tasks (with more detail in Section 7) and demonstrate that two variations of a standard diffusion model (Ratcliff, 1978; Ratcliff & McKoon, 2008) can account for these differences. Modeling results are compared with the only other model previously shown to be theoretically sensitive to both dimensions of the 2AFC task – The Leaky Competing Accumulator (LCA, Teodorescu, Moran, & Usher, 2016; Usher & McClelland, 2001).

* Corresponding author at: The Ohio State University, 291 Psychology Building, 1835 Neil Avenue, Columbus, OH 43210, United States.
E-mail address: ratcliff.22@osu.edu (R. Ratcliff).

https://doi.org/10.1016/j.cogpsych.2018.02.002
Accepted 11 February 2018
Available online 23 March 2018

When subjects are shown a single stimulus and asked to decide which of two categories it belongs to (e.g., is this a word or not, does this patch of pixels contain more black or more white pixels), evidence for the stimulus belonging to one of the categories is necessarily evidence against it belonging to the alternative category. In contrast, in the 2AFC task, when two stimuli are presented and the task is to decide which is more extreme on some scale, then strong evidence in favor of one stimulus is not necessarily evidence against the other stimulus (because the other stimulus could be strong or weak). Thus in 2AFC tasks there can be effects of both stimulus differences (the relative strength difference between the two presented stimuli) and overall stimulus magnitude (the total strength of the two items).

When there is a greater difference between the two stimuli (on the decision-relevant dimension) then subjects tend to respond more quickly and more accurately. This effect has been observed in a variety of domains (perceptual decision-making: Polania, Krajbich, Grueschow, & Ruff, 2014; Teodorescu & Usher, 2013; value-based decision-making: Hunt et al., 2012; Polania et al., 2014) and is analogous to difficulty effects observed in standard 2-choice tasks (Ratcliff & McKoon, 2008). However, the overall magnitude of the pair of presented stimuli also affects both RT and accuracy in the 2AFC task. This effect has been observed in a variety of tasks. Hunt et al. (2012) and Polania et al. (2014) found that more desirable pairs of stimuli produced shorter reaction times than less desirable pairs of stimuli in a value-based task. Teodorescu et al. (2016), Hunt et al. (2012), and Polania et al. (2014) found that larger or brighter pairs of stimuli produced shorter reaction times than smaller or darker pairs of stimuli in perceptual tasks, and Bowles and Glanzer (1983) found that pairs of old stimuli produced shorter reaction times than pairs of new stimuli in a memory task. In a numerosity discrimination task, Ratcliff and McKoon (in press) found that for a constant difference in numerosity between two arrays, as overall numerosity increased, accuracy decreased and reaction time decreased. Last, Niwa and Ditterich (2008) and Pirrone, Azab, Hayden, Stafford, and Marshall (2017) found speedups in RT with increasing stimulus magnitude for decisions with zero evidence (i.e. choice between N equal value alternatives).

2. Modeling

The overall magnitude effects provide a set of behavioral results that are not observed in standard 2-choice data that use a single 1D stimulus. As Teodorescu and Usher (2013) have pointed out, since these 1D stimuli are inherently competitive, standard 2-choice tasks are not ideal for distinguishing between models with different forms of competition (e.g., feed-forward input competition, as in some variants of the diffusion model, or response competition, as in LCA). In most sequential-sampling models for 2-choice decisions, information from the stimulus is accumulated noisily until some decision threshold is reached and then a response is made. There are many ways in which models of decision-making can differ (the nature of evidence accumulated, the competition between the two response alternatives, the stopping rule, etc), but the most relevant aspect in terms of fitting the overall magnitude effect in 2AFC data is the nature of evidence being accumulated.

There are several ways that decision models can represent the nature of evidence from the stimulus. In models with a single accumulation process (such as the diffusion model; Ratcliff & McKoon, 2008), evidence on each trial is typically modeled as a single value representing the rate of accumulation for that item. In a standard 2-choice task, this value would represent the item's strength on some dimension relevant to the decision (e.g., memory strength) relative to a neutral criterion value. In a 2AFC task, this value could represent the strength difference between the two presented items (again, on some relevant dimension). However, while difference based evidence representation would, naturally, be able to handle magnitude difference effects, it would be unable to handle overall magnitude effects that maintain a constant magnitude difference. This invariance prediction is a direct consequence of the fact that only the magnitude difference would be available to the decision process. In models with multiple accumulation processes (e.g., LCA, Usher & McClelland, 2001; LBA, Brown & Heathcote, 2008; accumulator models, Vickers, 1970), each accumulator receives input in favor of one of the response alternatives. In a standard 2-choice task, these inputs may be normalized to reflect the uni-dimensional nature of the stimuli (i.e., evidence in favor of one response is evidence against the other response). However, in a 2AFC task this normalization may not be appropriate and evidence in both accumulators might increase with stimulus magnitude.

On the other hand, accumulator models without competition may be able to handle overall magnitude effects, but have difficulty producing magnitude difference effects. For example, Hunt et al. (2012, supplementary information) used simulations of an independent race model to demonstrate that overall magnitude values produced large effects while magnitude differences produced fairly small effects. In contrast, accumulator models with competition can produce both patterns of effects. Teodorescu et al. (2016) demonstrated that competitive two alternative accumulator models could qualitatively handle both the overall magnitude and magnitude difference effects though they differed with how well they fit the data. They found that a version of the diffusion model with an accumulation rate based on the stimulus difference could account for both patterns of effects when the within-trial noise in the accumulation process was dependent on the overall magnitude of the stimuli. They also found that the Leaky Competing Accumulator (LCA) model could account for both patterns of effects because of competition between accumulators from lateral inhibition. Although the input to the accumulators in the LCA model is absolute, the lateral inhibition between the accumulators allows the model to produce difference effects that evolve and grow over time. Since the inhibition is based on the amount of accumulated evidence, towards the beginning of each trial the amount of inhibition is small and the accumulation process is sensitive to magnitude effects, similar to that of an independent race model (see also Niwa & Ditterich, 2008, for a comprehensive discussion). As more evidence is accumulated, however, the amount of inhibition increases such that the accumulation process becomes more similar to that of a diffusion model (see Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006; Heathcote, 1998; Marshall et al., 2009). Teodorescu et al. (2016) examined these effects using a single task and three conditions: (1) a baseline condition; (2) a condition in which the difference between the pair of stimuli was constant and the overall brightness increased compared to baseline; and (3) a condition

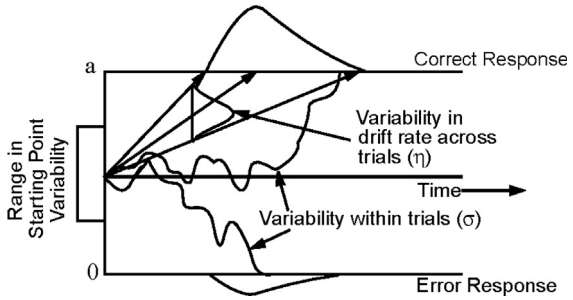

Fig. 1. Diffusion model. Evidence accumulation begins at starting point (a/2) and continues until one of the boundaries (0 and a) is reached.

where the ratio between the pair of stimuli was constant and the overall brightness increased compared to baseline. While these three conditions were sufficient to rule out several other models, they were not sufficient to distinguish between these LCA and diffusion models.

Our aims are first, to provide a more thorough examination of these overall magnitude effects by investigating multiple tasks with parametric manipulations of stimulus differences and magnitudes leading to a greater range of accuracy values. The second aim is to use the data to provide the basis for stronger tests of the LCA and diffusion models. The third aim is to present a different version of the diffusion model that can also account for these patterns of effects.

2.1. Diffusion models

In the diffusion decision model (Ratcliff, 1978; Ratcliff & McKoon, 2008), information from the stimulus is accumulated noisily from a starting point z until one of the two response thresholds is reached, either 0 or a, following which, a response is made (see Fig. 1). The model assumes that processes outside the decision process are represented by a single value, the nondecision time (T_{er}) that has uniformly distribted trial-to-trial variability with range s_t. The starting point is also assumed to have uniformly distributed variability from trial to trial with range s_z. Detailed expressions for accuracy values and distributions of response times can be found in Ratcliff and Smith (2004; see also Ratcliff & Tuerlinckx, 2002). The amount of time taken to reach one of the thresholds determines the RT and which threshold is hit determines the accuracy of the response. In the model, evidence on each trial is typically modeled as a single value (called the drift rate, which varies from trial to trial) representing the rate of accumulation for that item. In a standard 2-choice task, this value would represent the item's strength on some dimension relevant to the decision (e.g., brightness) relative to a neutral criterion value. In a 2AFC task, this value could represent the strength difference between the two presented items (e.g., the difference in brightness). While this evidence representation is able to handle magnitude difference effects, it is unable to handle overall magnitude effects since only the magnitude *difference* is available to the decision process. However, either one of two simple modifications enables the model to account for both magnitude difference and overall magnitude effects. In the first diffusion model variant, for each pair i, drift rate, ν_i, is a linear function of the difference in magnitude (Eq. (1)) and the trial-to-trial standard deviation in drift rate, η_i, is a constant times the sum of the squares of the magnitudes (Eq. (2), where the S values are stimulus strengths and d_1 and e_1 are coefficients that provide the scaling from stimulus strength to drift and across-trial variability in drift, respectively). This model has been used in modeling numerosity judgments (Ratcliff & McKoon, in press) and the relationships between the patterns of data are discussed later.

$$\nu_i = (S_{i1} - S_{i2})d_1 \tag{1}$$

$$\eta_i = e_1\sqrt{S_{i1}^2 + S_{i2}^2} \tag{2}$$

An alternative diffusion model is one that has the same linear relationship between drift and stimulus magnitude difference (and across-trial variability in drift rates is a constant), but in which within-trial variability, σ_i, is a function of the overall magnitude of the stimuli as in Eq. (3) (Teodorescu et al., 2016; for a model in which σ varies over the time course of the decision, see Smith & Ratcliff, 2009). For this version of the model, the within-trial variability is fixed to 0.1 in one of the conditions (to serve as a scaling parameter) and is defined in Eq. (3) (where the S values are stimulus strengths and s_1 is a coefficient that provides the linear scaling from stimulus strength to standard deviation).

$$\sigma_i = 0.1 + (S_{i1} + S_{i2})s_1 \tag{3}$$

Although these two Eqs. (2) and (3) have somewhat different structures, for values of S_1 and S_2 over the range 30–70, a plot of the two functions against each other is almost linear.

2.2. LCA model

In the LCA model, information from the stimulus is fed into the system via a separate noisy channel for each response alternative. When a single stimulus is displayed these inputs are commonly normalized to capture the one dimensional character of the information. When separate stimuli represent separate choices the input I_i represents the magnitude of stimulus 'i' (a linear function

thereof), as shown in Eq. (5). Accumulation of information for response alternative 'i' occurs in accumulator X_i and is driven by the magnitude of the input from channel 'i' minus a proportion β of the sum of accumulated evidence for other response alternatives $j \neq i$ (lateral inhibition) minus a proportion γ of its own accumulated evidence (leakage), as shown in Eq. (4). Response time and choice are jointly determined by the first accumulator to cross a common threshold.

$$X_i(t+1) = (1-\gamma)*X_i(t) + I_i(t) - \beta \sum_{j \neq i} X_j(t) \tag{4}$$

$$I_i = S_i \tag{5}$$

3. Experiments

We present data from five 2AFC perceptual experiments with both stimulus differences and magnitudes of the stimuli manipulated. Results show that both versions of the diffusion model (the η model and the σ model) are able to account for difference and magnitude effects in these tasks.

The first three experiments consisted of brightness discrimination tasks in which subjects were presented with two arrays of black and white pixels on a black background and asked to judge which of the two arrays contained more white pixels (i.e., which was brighter). For the first two experiments, the two arrays were dynamic (i.e., on each frame of the display, at a 60 Hz rate, a different random assignment of the dots was presented) and for the third experiment the two arrays were static. We varied both the difference in brightness between the two arrays and the overall brightness level. The fourth experiment consisted of a motion discrimination task in which subjects were presented with a patch of moving dots that contained some proportion of dots moving coherently in each of two different directions (right and left). The task was to decide whether a greater proportion of dots were moving coherently to the right or to the left. We varied both the difference in coherent motion probability between the two motion directions and the overall probability of coherent motion in each direction. The fifth experiment consisted of a brightness discrimination task in which subjects were presented with two flickering grayscale patches and asked to judge which patch was brighter. As before, we manipulated both the difference in brightness between the two patches and the overall brightness. The methods for each experiment are presented below and then all results from all the experiments are presented simultaneously.

3.1. Experiment 1: Flickering pixel displays

3.1.1. Method

3.1.1.1. Subjects. 25 Ohio State University undergraduates took part in the experiment for course credit.

3.1.1.2. Materials. All experiments were run on computers running the Linux operating system with a customized real-time system. Computers were connected to 17-in. monitors with a resolution of 640 × 480 pixels and a standard 102-key keyboard. Stimuli for the experiment consisted of paired arrays of black and white pixels on a black background, as shown in Fig. 2A. The centers of the two

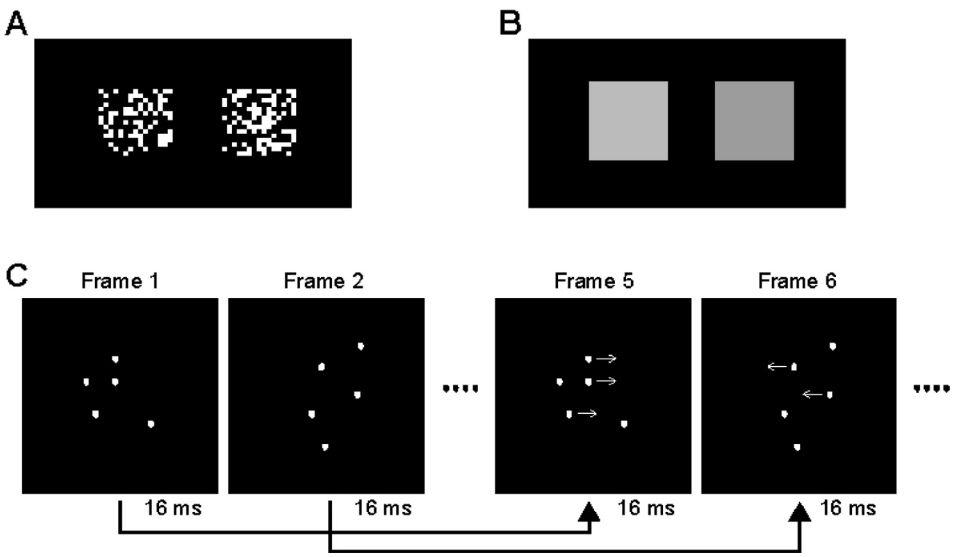

Fig. 2. Example stimuli. A: Example of the black and white pixel arrays for Experiments 1–3. For this pair, the array on the right contains more white pixels. B: Example of the grayscale patches for Experiment 5. For this frame, the patch on the left is a lighter shade of gray. C: Example of the sequence of images used in Experiment 4. Each frame is displayed for 16 ms. After the first four frames, each frame is an updated version of the frame presented 4 frames earlier. On updated frames, some dots are chosen probabilistically to move either right or left (on alternating frames) with the positions of the other dots changed randomly.

Table 1
Stimulus values for all experiments and conditions. Experiment 2 used the same values as Experiment 1. Values for Experiments 1–3 are the percentage of white pixels in each array of black and white pixels. Values for Experiment 4 are the probabilities of a dot moving coherently in the particular direction. Values for Experiment 5 are the grayscale values from a scale of 0 (black) to 255 (white). The grayscale values for Experiment 5 were converted to percentages when used in fitting the model.

Experiments 1 and 2		Experiment 3		Experiment 4		Experiment 5	
Array 1	Array 2	Array 1	Array 2	Motion 1	Motion 2	Array 1	Array 2
0.64	0.622	0.64	0.622	0.70	0.50	217	197
0.64	0.600	0.64	0.600	0.70	0.40	207	197
0.64	0.560			0.70	0.30	202	197
0.56	0.542	0.56	0.542	0.60	0.40		
0.56	0.520	0.56	0.520	0.60	0.30		
0.56	0.480			0.60	0.20		
0.48	0.462	0.48	0.462	0.50	0.30		
0.48	0.440	0.48	0.440	0.50	0.20		
0.48	0.400			0.50	0.10		
0.40	0.382	0.40	0.382	0.40	0.20	77	57
0.40	0.360	0.40	0.360	0.40	0.10	67	57
0.40	0.320			0.40	0.00	62	57

arrays were 100 pixels apart, and each array was 60 × 60 pixels. Each black or white patch within each array is a 4 × 4 patch of all white or all black pixels (i.e., the 60 × 60 pixel array consists of 15 × 15 patches of black or white pixels). At a standard viewing distance of 53.0 cm, each array was a square covering 3.24 degrees tall by 3.24 degrees wide with the two arrays covering 8.64 degrees from edge to edge. Both arrays flickered randomly at a rate of 60 Hz (i.e., approximately every 16 ms each array was replaced by a new randomly arranged array with the same percentage of white pixels). We measured the luminance values for different percentages of white pixels using a photometer and larger flickering pixel arrays and found a linear relationship between the percentage of white pixels in the array and the measured luminance.

3.1.1.3. Procedure. At the beginning of each trial, a white fixation cross was presented at the center of the screen for 600 ms. Subjects were told to focus their eyes on this cross. After the fixation cross disappeared, there was a delay of 200 ms and then the two stimulus arrays appeared. The stimuli remained on the screen until subjects made a response. Subjects were instructed to decide which of the two arrays was "brighter" (i.e., contained more white pixels). They made their responses using the 'z' and '/' keys on the keyboard with the 'z' response indicating they thought the array on the left was brighter and the '/' response indicating they thought the array on the right was brighter. If subjects responded incorrectly an error message was displayed for 750 ms. If subjects responded too quickly (RT shorter than 250 ms) a 'too fast' warning was displayed for 1500 ms. If subjects responded too slowly (RT longer than 1500 ms) a 'too slow' warning was displayed for 500 ms. Subjects completed 6 practice trials followed by 19 blocks of 72 trials each. We varied both the difference in brightness between the two arrays and the overall brightness level resulting in 12 conditions (3 difference values × 4 brightness levels). Stimulus values are displayed in Table 1. Each block of the experiment contained 6 trials from each of the 12 conditions, 3 with the brighter array on the right and 3 with the brighter array on the left all in random order. The ordering of trials within each block was otherwise random. The whole experiment took 45–60 min to complete.

3.2. Experiment 2: Briefly presented flickering pixel displays

3.2.1. Method

3.2.1.1. Subjects. 25 Ohio State University undergraduates took part in the experiment for course credit.

3.2.1.2. Materials and procedure. The materials and procedure were identical to Experiment 1, with the exception that stimuli were presented briefly (for only 300 ms) instead of remaining on-screen until subjects made a response. With a 300 ms stimulus presentation duration and new frames every 16.667 ms (60 Hz), this means that 18 frames were presented for every stimulus display.

3.3. Experiment 3: Static pixel displays

3.3.1. Method

3.3.1.1. Subjects. 16 Ohio State University undergraduates took part in the experiment for course credit.

3.3.1.2. Materials and procedure. The materials and procedure were identical to Experiment 1, with the exception that stimuli were presented briefly (for only 200 ms, i.e., 12 frames) instead of staying on-screen until subjects made a response, stimuli were static (i.e., were not replaced by a new pixel array every 16 ms), there were fewer conditions (see Table 1 for values), and there were 20 blocks with 96 trials per block.

3.4. Experiment 4: Motion discrimination

3.4.1. Method

3.4.1.1. Subjects. 26 Ohio State University undergraduates took part in the experiment for course credit.

3.4.1.2. Materials. Stimuli for the experiment consisted of a single 100 pixel diameter patch of moving dots. At a standard viewing distance of 53.0 cm, this patch was 5.4 degrees wide by 5.4 degrees tall. Each stimulus was composed of a series of frames displayed at a rate of 60 Hz. Each frame contained 5 dots (dots were single white pixels on a black background). On the first four frames the dots were located at random positions. On the fifth frame, a proportion of the dots from the first frame moved four pixels to the right and the other dots from the first frame moved randomly. On the sixth frame, a proportion of the dots from the second frame moved four pixels to the left and the other dots from the second frame moved randomly. This pattern continued with alternating frames containing dots moved to the right or left relative to the position of dots from four frames earlier, as shown in Fig. 2C. The proportion of dots moving in each direction was determined by the condition (see Table 1). Dot motion was determined in a probabilistic manner. The stimulus motion continued until subjects made a response.

3.4.1.3. Procedure. The stimuli were presented one at a time to the subjects and remained on the screen until subjects made a response. Subjects were instructed to decide whether a greater proportion of the dots were coherently moving to the right or to the left. Subjects made their responses using the 'z' and '/' keys on the keyboard with the 'z' response indicating they thought more dots were moving to the left and the '/' response indicating they thought more dots were moving to the right. If subjects responded incorrectly an error message was displayed for 500 ms. If subjects responded too quickly (RT shorter than 250 ms) a 'too fast' warning was displayed for 1500 ms. If subjects responded too slowly (RT longer than 1500 ms) a 'too slow' warning was displayed for 500 ms. Subjects were shown 8 example stimuli with various motion coherences followed by 19 blocks of 72 trials each. We varied both the difference in motion coherence between the two directions of motion and the overall motion strength resulting in 12 conditions (3 difference values × 4 coherence levels). Stimulus values are displayed in Table 1. Each block of the experiment contained 6 trials from each of the 12 conditions, 6 with the larger proportion of dots moving to the right and 6 with the larger proportion of dots moving to the left. The whole experiment took 45–60 min to complete.

3.5. Experiment 5: Grayscale brightness discrimination

3.5.1. Method

3.5.1.1. Subjects. 17 Ohio State University undergraduates took part in the experiment for course credit.

3.5.1.2. Materials. Stimuli for the experiment consisted of two homogenous, square, gray patches on a black background, as shown in Fig. 2B. The stimuli were presented on either side of a fixation-cross positioned at the center of the screen with the centers of the stimuli 100 pixels apart. At a standard viewing distance of 53.0 cm, each grayscale patch was 3.24 degrees tall by 3.24 degrees wide with the two arrays covering 8.64 degrees from edge to edge. For half of the trials the brighter patch was on the right, and for the other half of the trials the brighter patch was on the left. Each square's gray level fluctuated randomly and independently of the other patches over the course of each trial at a rate of 60 Hz. Brightness values were drawn from truncated normal distributions with mean levels presented in Table 1 (on a 0 (black) to 255 (white) scale). The distributions had standard deviation 25.5 and were truncated at ± 1 standard deviation to avoid obvious extreme values.

3.5.1.3. Procedure. At the beginning of each trial, a white fixation cross was presented at the center of the screen for 600 ms. Subjects were told to focus their eyes on this cross. After the fixation cross disappeared, there was a delay of 200 ms and then the two gray patches appeared. The stimuli remained on the screen until subjects made a response. Subjects were instructed to decide which of the two patches was "brighter" (i.e., was, on average, a lighter shade of gray). They made their responses using the 'z' and '/' keys on the keyboard with the 'z' response indicating they thought the patch on the left was brighter and the '/' response indicating they thought the patch on the right was brighter. If subjects responded incorrectly an error message was displayed on for 500 ms. If subjects responded too quickly (faster than 250 ms) a 'too fast' warning was displayed for 1500 ms. If subjects responded too slowly (slower than 1500 ms) a 'too slow' warning was displayed for 500 ms. Subjects completed 4 practice trials followed by 13 blocks of 72 trials each. We varied both the difference in brightness between the two arrays and the overall brightness level resulting in 6 conditions (3 difference values × 2 brightness levels). Stimulus values are displayed in Table 1. Each block of the experiment contained 12 trials from each of the 6 conditions, 6 with the brighter array on the right and 6 with the brighter array on the left. The whole experiment took 45–60 min to complete.

4. Behavioral results

For all of the experiments and analyses, responses were collapsed across right and left stimuli (or in the dot motion case, right and left motion). That is, correct responses to stimuli in which the brighter patch was on the right (or the stronger direction of motion was to the right) were combined with correct responses to stimuli in which the brighter patch was on the left (or the stronger direction of motion was to the left) and error responses from these same conditions were similarly combined. Response proportions and the 0.1, 0.3, 0.5, 0.7 and 0.9 response time quantiles for each condition and experiment are plotted in Fig. 3. For each of these plots, the

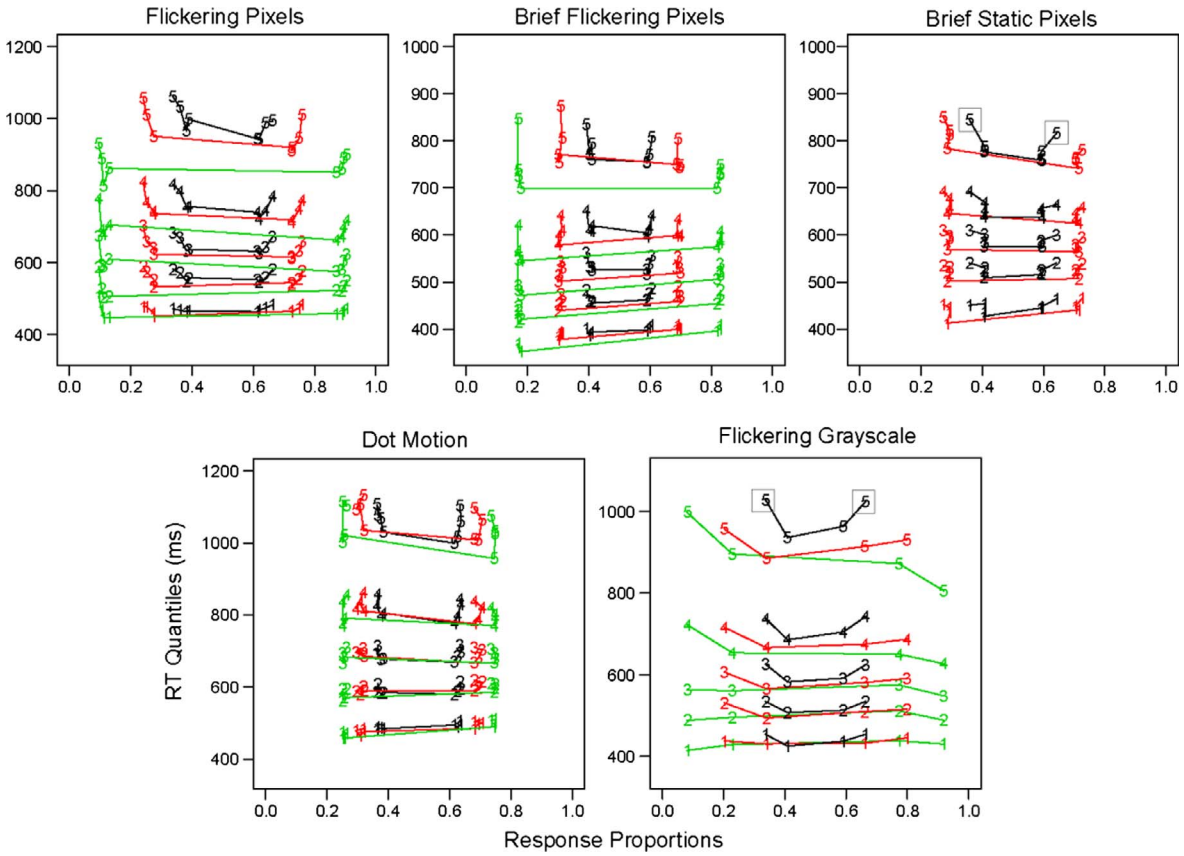

Fig. 3. Average data from all of the experiments. For each condition, response proportions are plotted along the x-axis and RT quantiles (0.1, 0.3, 0.5, 0.7, and 0.9) are plotted vertically for both correct and error responses (correct responses are those with response proportions greater than 0.5). The different colors are used to denote different difficulty levels (based on stimulus differences) with the easiest conditions plotted in green, medium conditions plotted in red, and hardest conditions plotted in black. (In this and the similar plots without color, green = light gray and red = dark gray). The different points along each line are the different overall magnitude conditions with the endpoints from each line coming from the lowest intensity conditions and the middle points coming from the highest intensity conditions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.) In this and other quantile plots, the horizontal lines are not predictions, they are there to show the conditions that correspond to each other.

conditions with the same stimulus difference are plotted in the same color (e.g., the conditions with largest differences between the pairs of stimuli are all plotted in green – the lines joining the left and right sides of the plot are for reference and do not represent predictions) with the endpoints for each line coming from the conditions with lower overall stimulus strength (i.e., darker stimuli or lower proportions of coherent motion). For example, in the plot of the data from Experiment 5 (Fig. 3: Flickering Grayscale), the two values in the gray boxes are the 0.9 RT quantiles for correct and error responses from the condition with grayscale values of 62 and 57. The other two points on that line are the 0.9 RT quantiles for correct and error responses from the condition with grayscale values of 202 and 197 (still a difference of 5, but overall greater brightness). Similarly, the two values in the gray boxes in the plot of the data from Experiment 3 (Fig. 3: Brief Static Pixels) are the 0.9 RT quantiles for correct and error responses from the condition with 0.40 and 0.36 percentages of white pixels and the other values along that line are from the other conditions with a 0.04 difference between the arrays of pixels (0.48/0.44; 0.56/0.52; 0.64/0.60 respectively). Across all of the tasks and most of the conditions, we see a pronounced u-shape to the data, with the conditions with lower overall stimulus intensity (the endpoints of the lines) having longer response times than the conditions with same difference but higher overall stimulus intensity (the other points along the lines). For most of these tasks, this pattern is observed for all conditions. However, for the flickering grayscale task this pattern is observed most strongly for the most difficult condition (i.e., the condition with the smallest difference between the two grayscale patches) but the shape flattens out for the correct responses for the easiest condition. This difference will be discussed later in the section on modeling. The effect of overall stimulus intensity on accuracy varies across tasks and conditions with some tasks showing almost no effect on accuracy (e.g., brief flickering pixels) and other tasks showing large effects on accuracy (e.g., flickering grayscale). Across all of the tasks we also see consistent difference effects, with largest differences between pairs of items (shown in green) producing higher accuracy values and shorter response times than intermediate differences (shown in red) and smallest differences between pairs of items (shown in black) producing the longest RTs and lowest accuracy values.

Table 2
Means (first lines) and standard deviations (second lines) for parameters from each experiment for the η model. a is the boundary separation, T_{er} is the mean non-decision time (sec), e_1 is the scaling parameter for across-trial variability in drift (Eq. (2)), s_z is the range of the across-trial variability of the starting point, p_z is the proportion of uniformly distributed contaminant responses (Ratcliff & Tuerlinckx, 2002 for details and parameter recovery), s_t is the range of the across-trial variability in non-decision time (sec), d_1 is the scaling parameter for drift rate (Eq. (1)), and p is the exponent in the transformation function used for Experiment 5.

Expt.	a	T_{er}	e_1	s_z	p_z	s_t	d_1	p
1	0.124	0.444	0.264	0.078	0.001	0.214	3.647	–
	0.014	0.049	0.136	0.039	0.000	0.064	1.202	–
2	0.104	0.395	0.246	0.075	0.001	0.195	3.153	–
	0.021	0.051	0.137	0.028	0.000	0.080	0.900	–
3	0.094	0.468	0.423	0.058	0.001	0.259	4.884	–
	0.016	0.049	0.160	0.031	0.000	0.060	1.577	–
4	0.129	0.488	0.285	0.099	0.001	0.319	0.371	–
	0.024	0.058	0.153	0.024	0.000	0.089	0.158	–
5	0.119	0.409	0.259	0.069	0.007	0.189	3.546	–
	0.016	0.046	0.158	0.045	0.024	0.055	1.311	–
5-P	0.123	0.410	0.196	0.072	0.008	0.193	5.616	0.577
	0.027	0.045	0.152	0.050	0.024	0.058	3.141	0.343

5. Model fitting results

We fit both versions of the diffusion model, the η model and the σ model, as well as the LCA model, to each individual subject's response proportions and reaction time quantiles (0.1, 0.3, 0.5, 0.7, 0.9) for each of the conditions. To fit the models, generic initial parameter values were chosen and then a simplex function (Nelder & Mead, 1965) was used to adjust the parameters of the model until the predictions matched the data as closely as possible. The match between the empirical data and the model predictions was quantified by a chi-square (χ^2) statistic, which was minimized by the simplex function (see Ratcliff & Tuerlinckx, 2002 for more detail). The RT quantiles divide the response proportion data into six bins each for correct and error responses, which gives 11 degrees of freedom per condition in the data. This gives 132 degrees of freedom for the experiments with 12 conditions, 88 degrees of freedom for the experiment with 8 conditions, and 66 degrees of freedom for the experiment with 6 conditions. The σ model has 9 free parameters and the η model has 8 free parameters. Because these models both constrain drift rates to be a function of the stimulus strengths, neither model has any parameters that are allowed to vary across conditions which means that these models are highly constrained (especially for the experiments with a larger number of conditions).

5.1. Diffusion model fitting

When fitting the diffusion models, the starting point of the accumulation process was fixed to half of the boundary separation (i.e., equidistant between the 'correct' and 'error' response boundaries) because we collapsed the data across right and left responses. Drift rates were all constrained to be a linear function of the stimulus values (shown in Table 1). Mean parameter values and standard deviations for each diffusion model are shown in Tables 2 and 3 and goodness-of-fit measures are shown in Table 4. In general, the

Table 3
Means and standard deviations for parameters from each experiment for the σ model. a is the boundary separation, T_{er} is the mean non-decision time (sec), η is the standard deviation for across-trial variability in drift, s_z is the range of the across-trial variability of the starting point, p_z is the proportion of uniformly distributed contaminant responses, s_t is the range of the across-trial variability in non-decision time (sec), d_1 is the scaling parameter for drift rate, s_1 is the scaling parameter for within-trial variability (Eq. (3)) and p is the exponent in the transformation function used for Experiment 5.

Expt.	a	T_{er}	η	s_z	p_z	s_t	d_1	s_1	p
1	0.152	0.450	0.189	0.104	0.001	0.213	4.316	0.027	–
	0.039	0.051	0.108	0.042	0.000	0.062	1.756	0.022	–
2	0.118	0.395	0.187	0.087	0.001	0.195	3.74	0.015	–
	0.033	0.051	0.116	0.034	0.000	0.080	1.561	0.022	–
3	0.107	0.462	0.261	0.061	0.001	0.246	4.989	0.019	–
	0.024	0.046	0.115	0.041	0.000	0.046	1.681	0.013	–
4	0.152	0.498	0.227	0.119	0.001	0.325	0.507	0.026	–
	0.037	0.072	0.153	0.032	0.000	0.088	0.340	0.026	–
5	0.128	0.404	0.175	0.071	0.009	0.181	3.347	0.010	–
	0.022	0.035	0.111	0.047	0.018	0.048	1.057	0.009	–
5-P	0.129	0.404	0.156	0.070	0.011	0.183	5.801	0.008	0.537
	0.023	0.034	0.117	0.049	0.021	0.049	3.104	0.008	0.318

Table 4
χ^2 values for the η and σ models. For each experiment, the critical χ^2 value is shown along with the mean χ^2 from fits to individual subjects, standard deviation across subjects, and the number of subjects (N) with observed χ^2 values less than the critical value. The degrees of freedom for the critical χ^2 values are 11*c-p (where c is the number of conditions in the experiment and p is the number of free parameters in the model). Without the power function on stimulus inputs, the η model has 7 parameters and the σ model has 8 parameters. With the power function, each model has one additional parameter.

Expt	σ model				η model			
	Critical value	Mean	SD	N < Crit	Critical value	Mean	SD	N < Crit
1	151.0	182.7	134.9	14/25	152.1	167.8	38.8	11/25
2	151.0	179.9	59.6	8/25	152.1	186.3	63.4	9/25
3	101.9	149.8	113.1	5/16	103.0	141.4	82.5	6/16
4	151.0	158.9	27.6	13/26	152.1	182.6	74.4	10/26
5	76.8	108.7	25.6	3/17	77.9	103.5	27.8	3/17
5-P	75.6	83.2	18.6	7/17	76.8	84.5	16.8	6/17

diffusion models fit about as well as each other. All of the mean χ^2 values are greater than the critical values, however out of 109 individual datasets, the σ model was able to produce simulated data that were not significantly different from the observed data for 43 of the subjects and the η model for 39 subjects. For the η model, the e_1 coefficient parameters yield η values that range, for the most extreme conditions and values across all experiments, from 0.086 to 0.378. The range of values within each experiment is smaller (e.g., values range from 0.228 to 0.378 for Experiment 3 and 0.086 to 0.298 for Experiment 5). For the σ model, the s_1 coefficient parameters yield σ values that range, for the most extreme conditions and values across all experiments, from 0.100 to 0.134.

Averaged data and model predictions for the first four experiments and each model are shown in Fig. 4 and the data and model predictions for the fifth experiment are shown in Fig. 5A. In these figures, the numbers are the data and the lines and circles are the model predictions. All other plotting conventions are the same as Fig. 3. Note that both of the models are able to produce the u-shaped functions that are observed in the data, and both of the models provide a reasonably close fit to the data. The σ model produces u-shaped functions with smaller changes in accuracy (i.e., straighter sides as opposed to more angled sides) where the η model produces u-shaped functions with larger changes in accuracy. Each of these patterns is also observed in the data in different experiments – in the dot motion task, there were essentially no changes in accuracy across different levels of overall motion strength, whereas in the flickering grayscale task there were large changes in accuracy across different levels of overall brightness. The σ model produces u-shaped functions across all difficulty levels while the η model can produce N-shaped functions for conditions with higher drift rates. For example, in the flickering pixel and flickering grayscale tasks, the η model predictions for the easiest conditions (in green) produce opposite RT effects for correct and error responses (i.e., predict faster correct responses but slower errors with increasing stimulus intensity).

For all of the pixels tasks, there is a linear relationship between the percentage of white pixels in the stimulus and the luminance of the screen (as measured by a photometer). For the fits presented in Table 4, we assume that drift rate is linearly related to stimulus intensity and so drift rate is linearly related to the proportion of white pixels. It is well known, however, that perceived brightness of a stimulus is not the same as objective luminance. Typically, perceived brightness follows a power law so that perceived difference between two luminance values at high luminance is smaller than the same luminance difference at small values of luminance. In Experiments 1–3, we used objective luminance values transformed to drift rates in modeling behavior. However, it may be that fits are better if perceive luminance values are used.

To address this, we assumed that the expressions to produce drift rate had luminance values raised to a power and the power was an additional parameter of the model. The result was that the model was unable to converge in most of the cases to produce an estimate of the exponent that was in a reasonable range (e.g., 0.3–1.5 for example). We placed limits on the value of the exponent, for example, 0.3 and 1.5, and for many of the subjects, the exponent moved to one of those values.

This means that the major determinant of the quality of the fit is not the linearity or nonlinearity of the transformation from luminance to perceived brightness, rather other aspects of the data. To take this one step further, we fixed the exponent of the power function to 0.5 (Geisler, 1989) and refit the data from Experiments 1, 2, and 3. Results showed average chi-square values that were a little lower or a little higher than the averages in Table 4. For the η model, chi-squares for Experiments 1, 2, and 3 were 175.4, 189.0, and 145.8 and for the σ model, the values were 167.5, 183.3, and 143.5.

The main explanation for this lack of sensitivity is the limited range of luminances used in the experiments. For Experiments 1, 2, and 3, the maximum and minimum values of luminance were 52.7 and 27.0 cd/m^2. The luminance values were linear with the proportion of white versus black pixels, so the other luminances can be computed from the proportions in Table 1. In Fig. 5B, this corresponds to grayscale values of about 0.60–0.84 and the function in the range is approximately linear. In fact, the difference in luminance values from the brightest pairs with largest difference (0.64 and 0.56, Table 1, with luminances 52.7 and 46.3 cd/m^2) and the darker pairs (0.48 and 0.40 with luminances 33.5 and 27.0 cd/m^2) were 0.59 and 0.46 (the difference between the luminance values to the power of one half).

For the grayscale task (Exp. 5), however, the screens were not linearized (gamma corrected) to produce grayscale values that were linearly related to the screen luminance. The left side of Fig. 5B plots the relationship between the grayscale value (adjusted to a 0–1 scale) and the measured brightness of the screen (using a photometer). Unsurprisingly, given that the monitors were not gamma corrected, there is a non-linear relationship between these values. Stimulus values from two of the conditions (one brighter pair and

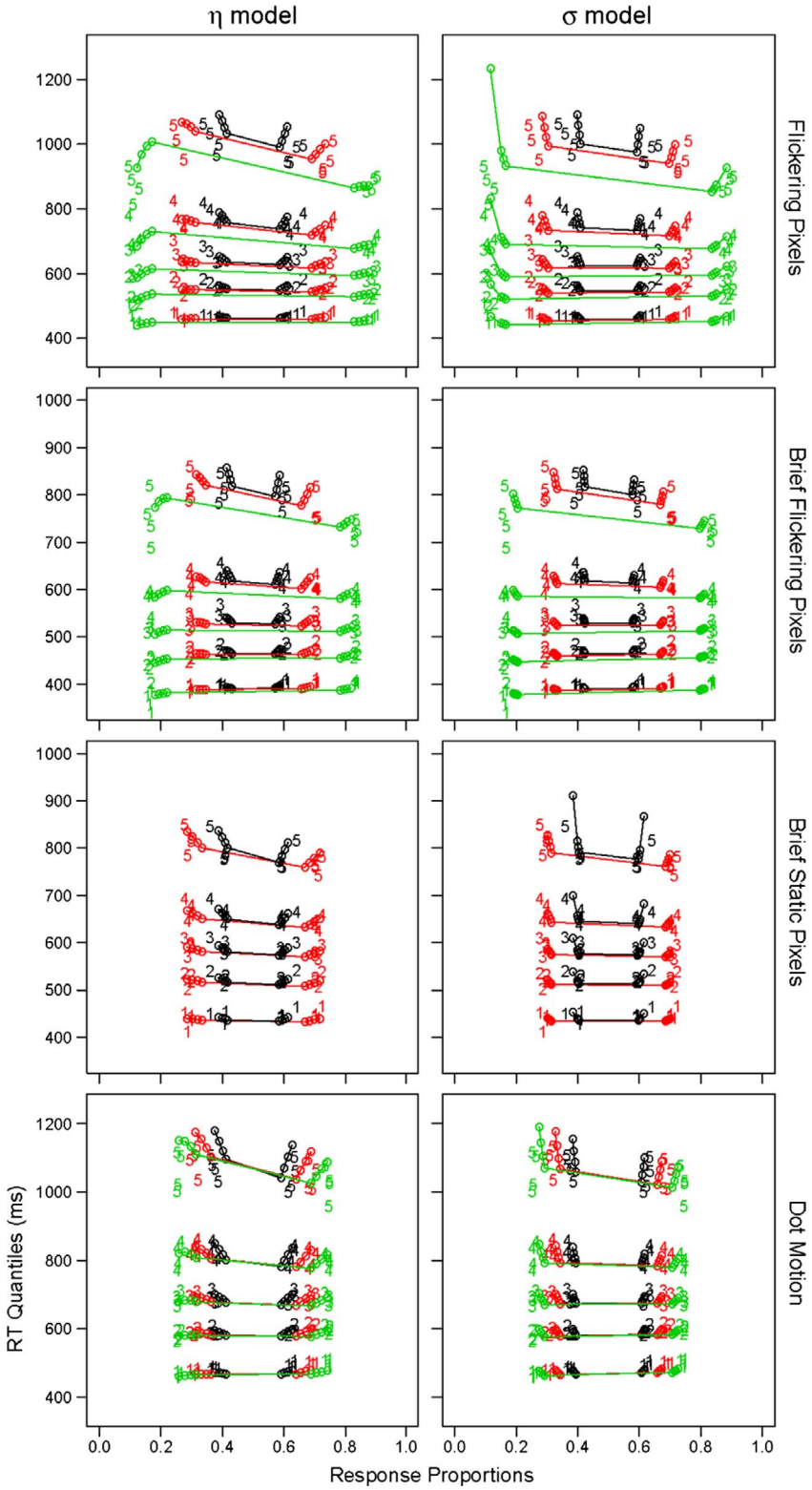

Fig. 4. Average data from the first four experiments and model predictions from the η and σ models. The same plotting conventions are used as for Fig. 3. The numbers are the behavioral data and the circles and lines are the model predictions.

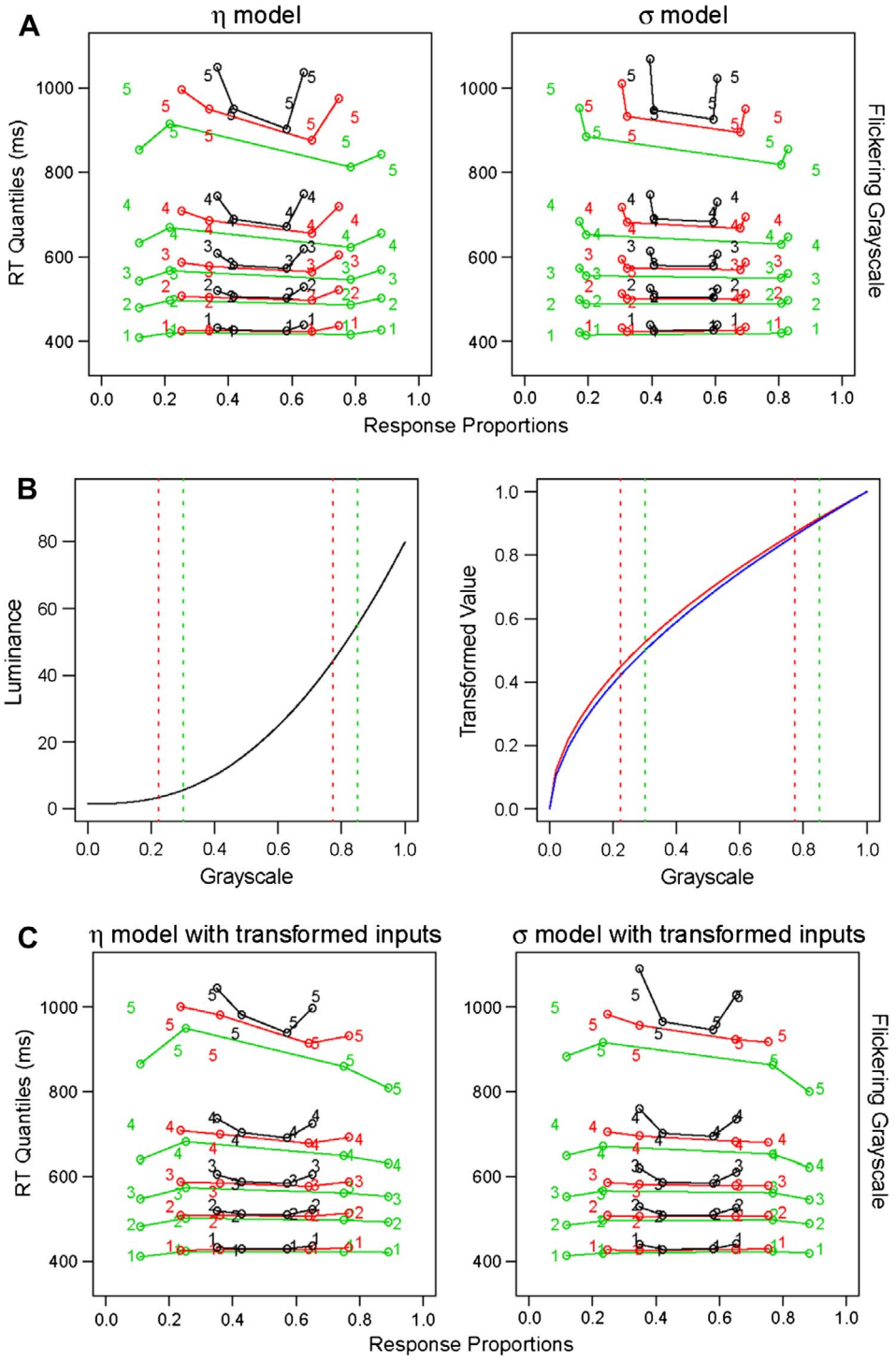

Fig. 5. Average data and model predictions for the flickering grayscale data. A: Fits of the models with untransformed grayscale values as inputs. The same plotting conventions are used as for Fig. 4. B: The figure on the left plots the screen luminance (as measured by a photometer) against the grayscale value (converted to a 0 to 1 scale). The figure on the right plots the transformed values against the grayscale values. The red line shows the transformed function for p = 0.537 (the average value from the σ model) and the blue line shows the transformed function for p = 0.577 (the average value from the η model). The dashed vertical lines in both figures indicate the grayscale values for the two conditions from the experiment with the most extreme grayscale values. C: Fits of the models with a power function transformation on the stimulus strengths. The same plotting conventions are used as for Fig. 4. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

one darker pair) are plotted as vertical lines. Although there was a constant difference between the stimuli in each pair on the input scale (the grayscale value), there was a larger luminance difference between the brighter pair than the darker pair. However, subjects were actually less accurate (on average) when presented with the brighter pair compared to the darker pair. Moreover, as mentioned above in the context of Exp. 1–3, even when the screen luminance is known to vary linearly, these values may be transformed when encoded by the perceptual system such that the resulting internal brightness representation is not linearly related to the physical value. These issues may contribute to the models' misses for the grayscale task.

To explore possible transformation between stimulus intensity (luminance) and the perceptual representation (brightness), we refit the data from the grayscale task (but not the pixel tasks) with grayscale values (from the computer grayscale range 0–255 divided by 255) transformed to luminance values ($S_{external}$) via a transformation that was determined by fitting a power-law function to the luminance values as determined by a photometer. Then the luminance values ($S_{external}$) were transformed into brightness values ($S_{internal}$) via a psychophysical power-law function (Eq. (6)). This required one additional free parameter p, as shown in Eq. (6).

$$S_{internal} = S_{external}^{p} \tag{6}$$

For these fits, these $S_{internal}$ values were then used in Eqs. (1)–(3) to generate mean drift rates and either across-trial or within-trial variability for each condition (depending on the model). The means and standard deviations for the parameter values for each model are shown in Tables 2 and 3 and χ^2 values are shown in Table 4 (as Experiment 5-P). Unlike for Exp. 1–3, the addition of the power function on stimulus strengths did improve the fit of both of the models. The average estimated values for the power function parameter, p, were similar across the two models at around 0.55 and the estimates for each subject were strongly correlated across the two models ($r(15) = 0.98$ $p < 0.01$). These parameter values replicate known results from psychophysical literature where the power–law coefficient for brightness is regularly found to be around 0.5 (Geisler, 1989). This value produces a compressive transformation such that the internal representations for pairs of higher intensity stimuli are closer together than the internal representations for pairs of lower intensity stimuli. This allows the models to produce the accuracy results observed in the grayscale task because the mean drift rates are now sensitive to both the overall and relative brightness of the stimuli. The right side of Fig. 5B plots the transformed grayscale values against the original values for the average value of p from each model. The average data and model predictions for the η and σ models with transformed inputs are shown in Fig. 5C. Note that both models are now able to produce the crossover pattern observed for correct responses: for more difficult conditions (shown in black) responses to brighter stimuli are made more quickly and less accurately, but for easier conditions (shown in green) responses to brighter stimuli are made more slowly and less accurately. Both models are still missing the tails of the RT distributions for the error responses, but there are relatively few errors for some of these conditions such that those quantiles may not be very reliable (some subjects made zero errors in the easiest conditions such that these averages reflect data from only a subset of subjects).

For this transformation, we assume that some mean stimulus value is encoded from the stimulus and transformed according to a power function either in the encoding process or in the process of extracting the decision relevant information. In both models, we then assume that the transformed value may vary across trials according to a normal distribution (in the σ model the variability parameter of this distribution is constant across conditions while in the η model the variability parameter of the distribution is related to the stimulus magnitude). Alternatively, the noise across trials or across samples may come into play earlier in the process such that it would be more appropriate to apply the transformation to distributions of stimulus values. In that case, the resulting internal representation of stimulus strength across trials would not follow a normal distribution. However, for the range of values used in these experiments and the average transformation, these alternatives provide fairly similar outcomes and the diffusion model has been shown to be robust to moderate changes in distributional assumptions (Ratcliff, 2013). In Fig. 6, the top panel shows a series of possible normal distributions of stimulus strength. When these distributions are transformed according to a power function with $p = 0.5$, the distributions are as shown in the middle panel of Fig. 6 and there is a roughly linear relationship between the standard deviations of the original distributions and the standard deviations of the transformed distributions (as shown at the bottom of Fig. 6). This indicates that, for our modeling framework, we should get similar results whether the noise is present in the stimulus (or early processing stages) and then transformed or the noise is added later on as a normal distribution with variability that scales with the stimulus mean.

These modeling results demonstrate that the effects observed in these 2AFC tasks can come about in several ways. Sensitivity to overall value may be related to internal processing noise such that larger inputs to decision process are accompanied by higher variability in the noise in the accumulation process (i.e., activation dependent noise, as in the σ model). This interpretation is consistent with the notion that neural firing rates are distributed according to a Poisson distribution. A characteristic property of Poisson distributions is that the variance is proportional to the rate such that higher firing rates should be accompanied by higher variability. Alternatively, stimuli with greater intensity may be processed with more variability across trials (as in the η model). For the experiments where the physical properties of the stimuli were known to follow a linear relationship with the experimental manipulations (i.e., the brightness tasks with black and white pixels), transforming the stimulus values via a psychophysical power-law transformation resulted in difficulties to converge while fixing the power coefficient to a psychophysically plausible 0.5 did not improve average chi-square values noticeably. For other stimuli (i.e., the grayscale patches) or other ranges of stimulus values, however, it may be necessary to transform the stimulus strengths to reflect that these stimuli may not be represented in the perceptual system by a linear scale.

6. LCA model

Teodorescu and Usher (2013) performed a series of forced-choice perceptual experiments aimed at distinguishing between

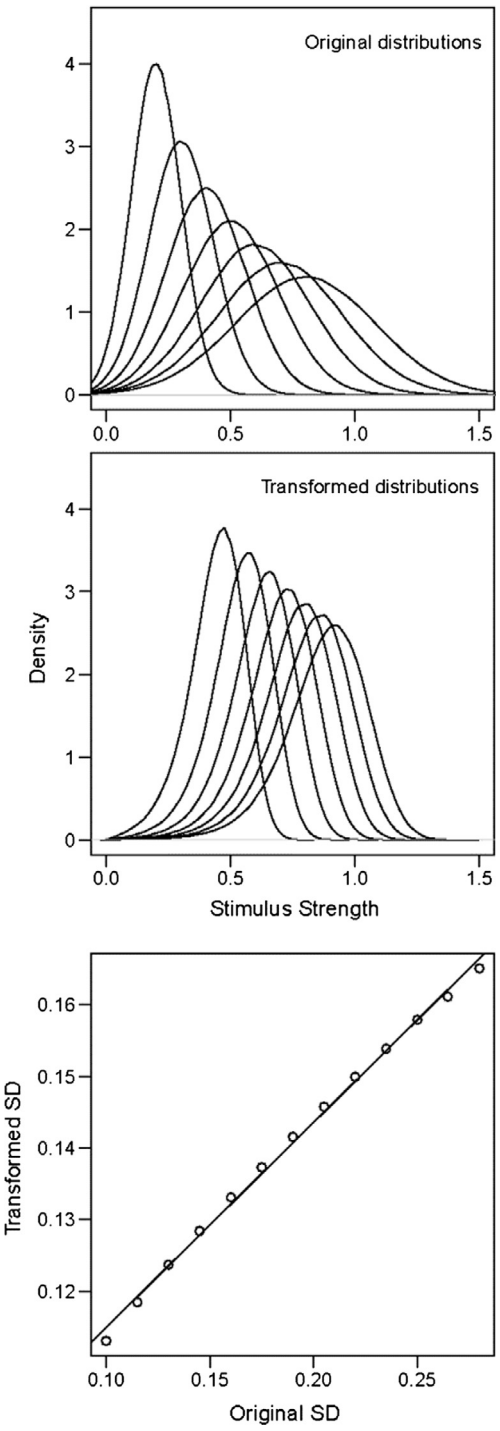

Fig. 6. Transformations of normal distributions of stimulus strength. The top figure shows hypothetical normal distributions of stimulus strength, with variance increasing with stimulus strength. The middle figure shows the square root of these values (e.g., power function with $p = 0.5$). The bottom figure shows the relationship between the standard deviations of the original distributions and the standard deviations of the transformed distributions along with the best-fitting line.

different forms of competition in decision models. In their Experiment 3b, subjects had to choose which of two flickering grayscale patches was brighter. On some of the trials, after 100 ms the brightness of both of the patches increased by a factor of 1.5 and remained at that level until the subject made a response. Relative to the responses in the neutral condition (in which the stimulus strength did not change during the trial), reaction times on these trials were shorter while accuracy remained the same (however, in their Experiment 3a there was a slight decrease in accuracy in this condition). However, it is important to note that multiplying by a

factor produces a mixed difference plus magnitude effect because the multiplicative operation magnifies the difference in addition to the overall values.

Subsequently, Teodorescu et al. (2016) conducted two experiments demonstrating people's sensitivity to the overall magnitude of stimuli and investigated which model architectures were able to account for the observed effects. In the experiments, subjects were asked to determine which of two patches of flickering grayscale values was, on average, brighter. Compared to a baseline condition, Teodorescu et al. increased the brightness of both of the stimuli such that either the difference in the mean brightness of the two patches remained constant (the additive condition) or the ratio of the mean brightness of the two patches remained constant (the multiplicative condition). Unlike the additive condition, in the multiplicative condition, both overall magnitude *and* difference were increased compared to baseline. In both the additive and multiplicative conditions, subjects made responses more quickly than in the baseline condition. Subjects were equally accurate in the baseline and multiplicative conditions and less accurate in the additive condition.

As described earlier, Teodorescu et al. (2016) were able to fit their results with two different models: a version of the diffusion model similar to the σ model used here, and a version of the Leaky Competing Accumulator (LCA) model. In both of their experiments, stimulus values varied over time according to a normal distribution (similar to our flickering grayscale experiment). In their model fits, samples from these distributions were transformed according to a power function and then used as inputs to the model accumulators. For both models, within-trial noise was allowed to include both a general component (constant across conditions) and a stimulus-dependent component that was determined by the transformed stimulus samples. For the diffusion model, the stimulus-dependent noise component enabled the model to account for overall magnitude effects. For the LCA model, however, the stimulus-dependent component went to zero indicating there was no added benefit in quality of fit for a relationship between the stimulus value and the internal noise for this model. Rather, the LCA was able to produce overall magnitude effects because of the dynamic nature of the evidence accumulation process. In the LCA model, evidence is accumulated in a stochastic fashion in two competing accumulators. The amount of evidence in each accumulator is affected by both inhibition between the two accumulators and decay (or leakage) within each accumulator, both of which are proportional to the total accumulated evidence. The inhibition term has less of an effect early on in the accumulation process (because less total evidence has been accumulated) such that in the early stages of accumulation the behavior of this model is more similar to that of an independent race model. Over time, however, the behavior of the model becomes more similar to that of a standard diffusion model as the amount of accumulated evidence increases such that both the inhibition between accumulators and leak increase (Bogacz et al., 2006). Stimuli with greater overall magnitude then primarily influence the earlier stage of this model such that the process more quickly approaches the threshold and more quickly transitions to the phase where it is more sensitive to the stimulus differences (i.e., similar to a diffusion model accumulating relative information but with lower threshold).

We fit an LCA model to our datasets to compare this model's performance with the versions of the diffusion model presented here. To keep the fit consistent with the diffusion models, we used the same scaling functions and restrictions whenever possible. That is, the inputs to the two accumulators in LCA were psychophysical power-law functions of the two stimulus strengths (screen luminance values as defined by $S_{external}$ in Eq. (6)) as shown in Eq. (7), where the drift rate for condition i and accumulator j is a psychophysical power-law function of the stimulus strength for that particular condition and response option (S_{ij}).

$$v_{ij} = (S_{ij}^{p})d_1 \tag{7}$$

Consistent with Teodorescu et al. (2016) we set $d_1 = 1$ to act as a scaling parameter, included a psychophysical power-law coefficient parameter p for all the fits, restricted the within-trial noise in the accumulation process to be constant across conditions, and did not include across-trial noise in drift. Mean parameter values and standard deviations for the LCA fits can be found in Table 5 and mean χ^2 values and standard deviations can be found in Table 6.

Overall, this model had difficulty producing the range of accuracy values observed in the experiments with a single set of

Table 5
Means (first lines) and standard deviations (second lines) are shown for parameters from each experiment for the LCA model. a is the height of the response boundary above zero divided by 100, s is the standard deviation of the Gaussian internal processing noise, T_{er} is the mean non-decision time in sec., s_z is the range [0, s_z] of uniformly distributed across-trial variability of the starting point dividied by 10, s_t is the range of the across-trial variability in non-decision time in sec., Time-step is the discrete time step of each simulation iteration in ms/100, and p is the exponent in the psychophysical power-law transformation function.

Expt.	Leak	Inhibition	a	s	s_z	T_{er}	Time-step	p	s_t
1	0.889	0.158	0.108	0.286	0.507	0.277	0.090	1.33	211
	0.009	0.006	0.005	0.022	0.036	0.034	0.005	0.08	19
2	0.916	0.156	0.165	0.577	0.521	0.189	0.104	1.524	167
	0.032	0.059	0.041	0.244	0.239	0.042	0.050	0.454	74
3	0.950	0.145	0.248	0.548	0.546	0.186	0.105	1.760	254
	0.042	0.066	0.087	0.342	0.351	0.067	0.075	0.475	71
4	0.929	0.139	0.106	0.454	0.684	0.275	0.185	0.529	226
	0.025	0.055	0.025	0.173	0.293	0.062	0.121	0.211	50
5	0.982	0.158	0.058	0.008	0.211	0.180	0.151	0.261	379
	0.112	0.052	0.017	0.030	0.031	0.048	0.023	0.299	276

Table 6
χ^2 values for the LCA model are shown. For each experiment, the critical χ^2 value is shown along with the mean χ^2 from fits to individual subjects, standard deviation across subjects, and the number of subjects (N) with observed χ^2 values less than the critical value. The degrees of freedom for the critical χ^2 values are 11*c-p (where c is the number of conditions in the experiment and p = 9 is the number of free parameters in the model).

Expt.	Critical value	Mean	SD	N < Crit
1	149.9	262.9	87.3	0/25
2	149.9	258.4	74.3	0/25
3	100.7	161.1	81.1	1/16
4	149.9	157.0	24.8	11/25
5	75.6	493.0	291.2	0/17

parameters (see Fig. 7). Correspondingly, the model performed best when there was a more limited range of accuracies as in Experiments 3 & 4 and performed poorly when the range of accuracies was more expansive as in Experiments 1, 2 & 5. While the model can produce the u-shaped functions and is in the right range in terms of RT predictions, it underestimates accuracy by various degrees for all conditions. Similar to the σ model, for Experiments 1 & 2, the LCA predictions show no difference in accuracy across different brightness levels (just a speed up in response times). Although the behavioral effects are small, this seems more in line with Experiment 2 in which stimulus strength has no effect on accuracy. For Experiment 3 the LCA model even predicts increased accuracy together with reduced RT for higher stimulus strength. Again, although the behavioral effects are small this seems to match the data. For Experiment 4 the underestimation of accuracies is the smallest with good fits to the data. For Experiment 5 the underestimation of accuracies is the largest which, together with considerable misses in RT, produced poor fits to the data.

For Experiments 1, 2 and 5, the LCA model had difficulty producing appropriate accuracy and response time values for the entire range of conditions within each experiment with a single set of parameters. That is, the inhibition, decay, and boundary values that could produce reasonable response time estimates when the stimulus values were larger (e.g., around 64%) did not produce reasonable response times when the stimulus values were smaller (e.g., around 32%) and vice versa. In some cases, the parameter values that worked well for one range of stimulus values could not even be used to generate predictions for other ranges of stimulus values because they produced accumulation processes that almost never terminated (e.g., a larger decay term would work with larger drift rates but not smaller drift rates). Similarly, the way the stimulus strengths are used to determine the accumulator's drift rates made it more difficult for the model to produce a wider range of accuracy values. While the LCA model is capable of producing a wider range of accuracy values and reasonable-looking RT distributions, for Experiments 1, 2 and 5, the predictions do not match the data in terms of which stimulus strengths correspond to which accuracy values.

It is also interesting to note that for Experiments 1, 2 and 3, while the fits were not very good, the psychophysical power-law coefficient parameter that fit the data best was substantially larger than one (~1.3, ~1.5 and ~1.7 for Experiments 1, 2 and 3 respectively). This would indicate an expansive (convex) psychophysical function for brightness. This is in contrast to, and inconsistent with established findings of coefficients circa 0.5, placing brightness in the range of compacting (concave) psychophysical functions. Because the LCA parameters are not always recoverable, even though we have fit the model several times, it is possible that other parameterizations might exist that give similar fits with a more conventional psychophysical function. For Experiment 4 with random dot motion, the LCA provided good fits with reasonable power-law coefficients around 0.5.

If the drift rates were allowed to vary freely, then the LCA model might be able to produce a wider range of accuracy values for these tasks. Adjustments to other model parameters (i.e., within-trial noise, boundary height, inhibition, and decay parameters) can also enable the model to produce a wider range of accuracy values. Note however, that the LCA model is technically more difficult to fit due to non-linearities and complex interactions between the leak and inhibition parameters, and has been shown to have problems in parameter recovery (see Miletić, Turner, Forstmann, & van Maanen, 2017, although their application is to a 3-choice version of the model with a different drift rate parameterization). So although we were unable to fit some of these data with the LCA model, we do not feel this is the last word and other versions of the LCA might not fail.

7. Discussion

We observed effects of both overall stimulus magnitude and stimulus difference effects in several perceptual tasks across a range of difficulties. In these tasks, subjects responded more quickly and more accurately when there was a larger difference between the pair of stimuli, and subjects also responded more quickly when the overall magnitude of the pair of stimuli was greater. These results help validate the effect of overall magnitude on decision-making, which has previously been sporadically observed in a handful of tasks and often with only a small number of conditions. We demonstrated that two highly constrained versions of the diffusion model were able to account for both patterns of effects in these tasks. In both versions of the diffusion model, drift rates were constrained to be a linear function of the difference between the pairs of stimuli (or the power-law transformed differences in Exp. 5). In the σ model, the within-trial noise parameter increased linearly with the overall stimulus magnitude. In the η model, the across-trial noise parameter increased as a function of the square root of the sum of the squares of the two stimulus magnitudes (i.e., independent sources of noise). Both of these models were able to produce both stimulus difference and overall magnitude effects consistent with the data. However, some stimuli required additional transformations on stimulus strength to reflect how information is represented in the perceptual system. For the brightness tasks with black and white pixels and for the dot motion task, a linear mapping from stimulus intensity to drift rate was sufficient to produce the observed patterns of effects and a power law transformation did not

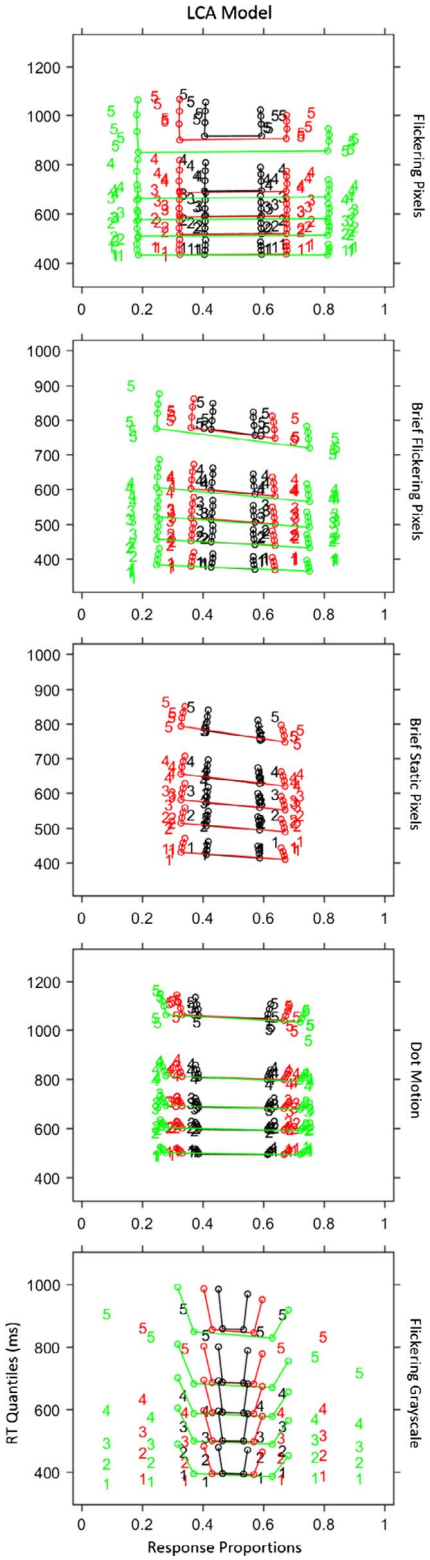

Fig. 7. Average data from Experiments 1–5 and model predictions from the LCA model. The same plotting conventions are used as for Fig. 3. The numbers are the behavioral data and the circles and lines are the model predictions.

improve the fits. For the brightness task with grayscale patches, a transformation of the stimulus values was necessary to enable the models to produce the crossover pattern observed for correct response-times. The LCA model with psychophysical transformation fit some of the data sets well but failed for most data sets. Specifically, the LCA performed best when the range of accuracies was small and performed worst when the range of accuracies was large.

7.1. Overall magnitude effects

The overall magnitude effects have been observed in several other tasks. Hunt et al. (2012) observed an overall magnitude effect in a value-based 2AFC task in which subjects had to choose between a pair of gambles with varying reward probabilities and reward amounts. Each gamble was assigned a subjective expected value (based on prospect theory) and the researchers examined the effects of these values on behavioral and MEG data. Behaviorally, Hunt et al. (2012) observed that both overall expected value (the sum of the subjective expected values of the pair of gambles) and the expected value difference had negative relationships with reaction time. Both larger value differences between the two gambles and larger overall values led to shorter reaction times. The effect of value difference, however, was much greater than the effect of overall value. In terms of MEG data, Hunt et al. (2012) found that overall value had an effect on activity in the 3–9 Hz range and value difference had an effect on activity in the 2–4.5 Hz range. The effect of overall value began earlier and had a slightly shorter duration than the effect of stimulus difference. The effect of overall value was significant for both correct and error responses whereas the effect of value difference was only significant for correct responses.

Polania et al. (2014) found an effect of overall stimulus magnitude in 2AFC tasks in both perceptual decision making and value-based decision making tasks. Polania et al. (2014) also found an effect of overall magnitude on accuracy, but this effect was not consistent across the perceptual and value-based tasks. In the perceptual task, there was a negative relationship between overall magnitude and accuracy (i.e., larger stimuli produced less accurate responses). In the value-based task, there was a positive relationship between overall magnitude and accuracy (i.e., more positively rated stimuli produced more accurate responses). There was also a difference across the two tasks in the relative size of the overall magnitude and magnitude difference effects. For value based decisions, the effect of overall magnitude was larger than the effect of magnitude difference. For perceptual decisions, it was the opposite (the effect of magnitude difference was greater). These differences may be a function of task differences and the motivation of the subjects. In the value-based task, subjects were choosing which of two food items they would prefer to eat. In the perceptual task, subjects were choosing which of two food items is larger. Intuitively, choosing between two food items that are liked may be qualitatively different than choosing which of two large items is bigger. In the food case, although there is technically a wrong answer based on subjects' previous ratings of the food items, the 'wrong' answer is more appealing when it's an option the subject likes. In contrast, in the perceptual task the overall size of the items shouldn't have any effect on subjects' motivation.

Data from Bowles and Glanzer (1983) could be interpreted as demonstrating an overall magnitude effect in a 2AFC memory task. In the experiment, subjects were shown pairs of words and asked to indicate which word had been in the previous list. Most of the pairs consisted of one old word and one new word, but some of the pairs consisted of either two new words or two old words. Overall, reaction times were longest for the pairs consisting of two new items and were considerably longer than reaction times for the pairs consisting of two old items (about 875 ms longer on average). Assuming that the memory strength differences between pairs of old items is comparable to the memory strength differences between pairs of new items, this difference in reaction time would be consistent with an overall magnitude effect. However, it is also possible that old items are more variable than new items such that it would not be reasonable to assume that the difference in memory strength for pairs of new items is comparable to the difference for pairs of old items (and so the observed behavior could be the result of a difference effect as opposed to an overall magnitude effect). Both signal detection modeling (Cohen, Rotello, & Macmillan, 2008; Heathcote, 2003; Hirshman & Hostetter, 2000; Mickes, Wixted, & Wais, 2007) and RT modeling (Starns, Ratcliff, & McKoon, 2012) results support the notion that old items are more variable than new items. Bowles and Glanzer (1983) also found accuracy results consistent with a standard SDT representation for high- and low-frequency words in memory such that low-frequency words had a larger d' and the memory strength distributions were ordered as follows (from weaker to stronger): low-frequency new items, high-frequency new, high-frequency old, low-frequency old. If this is an accurate representation of the relative memory strengths of the words in their experiment, then some of their reaction time results for various old/new pairings are consistent with an overall magnitude effect. For example, when subjects were presented with a low-frequency old word paired with a high-frequency new word (LO, HN pair), their mean response times were shorter than when presented with a low-frequency old word paired with a low-frequency new word (LO, LN pair). This result is consistent with an overall magnitude effect (the LO, HN pair has greater overall memory strength than the LO, LN pair) but inconsistent with a magnitude difference effect (there is greater separation between the two low-frequency distributions so the LO, LN pair should yield the smallest RTs). Other comparisons of word-pairs are, however, more consistent with a magnitude difference effect. Thus it is likely that both effects are present in these data, as in the value-based and perceptual tasks described previously. Bowles and Glanzer (1983) did fit their data using a stimulus-sampling model of recognition memory. While the model was primarily designed to capture certain accuracy and interference effects, their reaction data was handled reasonably well by the assumption that reaction time was a decreasing linear function of the difference between the two items (the "difference" between the items was determined by their model).

Ratcliff and McKoon (in press) found effects of both overall magnitude and stimulus differences in a variety of numerosity tasks. In some of their experiments, subjects were presented with an array of intermingled blue and yellow dots and asked whether there were more blue or more yellow dots. They varied both the total number of dots (magnitude) and the difference between the number of blue and yellow dots. Subjects were found to respond more quickly when there was a larger difference between the colors but also when

the total number of dots increased. Interestingly, they also found that the cognitive representation of numerosity and the effects of confounding variables were task-dependent. When subjects were asked to judge which of two side-by-side arrays contained more dots, they responded more slowly when the total number of dots increased. When Ratcliff and McKoon varied the area of the dots, they found that area had a larger effect on performance in the tasks where the dots were intermingled than in the tasks where the dots were presented side-by-side. Similar to our results, when the dots were intermingled, they were able to fit the overall magnitude and stimulus difference effects with a version of the diffusion model where the drift rate was a linear function of the stimulus difference and across-trial variability in drift increased with the overall stimulus magnitude. For the side-by-side arrays, they were able to fit the observed effects with a version of the diffusion model where the stimulus values were represented on a log scale and across-trial variability in drift was (mostly) constant across conditions. They hypothesized that when dots were intermingled, the processing system could not form separate representations of the different colors of dots and the linear model was preferred. But when the arrays were separated, separate representations of the two arrays could be produced leading to a preference for the log model. The parallel interpretation for the results presented here would be that for the perceptual tasks presented here, separate representations cannot be produced (cf. a failure of individuals to perform absolute identification tasks for such stimuli) and so the linear η model would be preferred over a log model (which would fail completely). The difference between the brightness tasks and numerosity tasks is that for numerosity, it is possible to guess the number of dots in an array, but for brightness, there is no simple label for the brightness of a patch. This may explain the difference between side by side dot arrays versus side by side pixel or gray scale arrays.

Simen, Vlasov, and Papadakis (2016) found simultaneous stimulus difference and overall magnitude effects in a brightness discrimination task, an auditory discrimination task, and a vibrotactile discrimination task. For each of their tasks, the high-intensity stimulus strengths were multiples of their low-intensity stimulus strengths such that the high-intensity stimulus pairs had both greater overall magnitude and larger differences between pairs (as in the multiplicative condition in Teodorescu et al., 2016). They demonstrated that a diffusion model that approximated a spike-counting process could account for the observed patterns of results. This model assumes that stimulus intensities are represented by proportional firing-rates in some neural population and these representations feed into a counter with a response threshold. Because these firing rates follow a Poisson process, as stimulus intensity increases, variability also increases such that in the diffusion model approximation of the Poisson process, stimulus intensity is related to within-trial accumulation noise (as in our σ model).

7.2. Stimulus representations

For most of our stimuli, we were able to obtain reasonable diffusion model fits by assuming a linear relationship between physical stimulus intensity and the internal stimulus representation, and allowing noise (whether internal or across trials) to increase with stimulus intensity. This is consistent with Thurstone's (1927) law of comparative judgments as it relates to Weber and Fechner's work on perception. In a series of discrimination experiments in the early 19th century, Weber demonstrated that the just-noticeable difference (JND) between two stimuli (Δx) is proportional to the absolute magnitude of the stimuli (x). This can be expressed as in the equation below, where 'k' represents the constant "Weber fraction" (this equation is also known as "Weber's Law):

$$k = \frac{\Delta x}{x}$$

Fechner (1860) assumed that this reported difference really reflects a JND in subjective sensation (Δs, which is proportional to k) as opposed to a physical difference. Based on Weber's law he derived a relation between the physical stimulus magnitude and the sensed magnitude. This equation is known as the "Weber-Fechner law" (where x_0 is the threshold for perception):

$$s \propto ln\frac{x}{x_0}$$

More recently, Stevens (1957) used a different kind of task (a matching task instead of a discrimination task) and demonstrated that the relationship between the magnitude reproduced by subjects (r) and the physical magnitude was best captured by a power function:

$$r \propto x^b$$

Later work has proposed that the two different laws may simply reflect different stages in the processing of magnitude (MacKay, 1963; Shepard, 1981).

From a processing point of view, the general decrease in perceptual sensitivity associated with increases in overall magnitude has been suggested to result from two alternative mechanisms (Geisler, 1989). One mechanism implies a non linear compressive function on internal stimulus magnitude representations (as in the formulation of Fechner's law) with constant internal noise being added after the transformation. Another mechanism assumes a linear relation between physical stimulus magnitude and internal magnitude representation but with a monotonic increase in internal noise with increased magnitude. While one or the other mechanisms is enough to generate the behavioral effect, it is possible that the two mechanisms are complementary and both play a causal role in producing diminishing sensitivity with increased magnitude.

It was also recognized that the stimulus intensity necessary to pass the threshold and be perceived varied from trial to trial in an apparently random manner such that multiple measurements were required to estimate the threshold value. According to Thurstone's theory of judgments (1927), the perception of stimulus properties is intrinsically variable and this variability is what gives rise to variability in people's judgments. This perceptual representation is assumed to be normally distributed. Although this general idea

was not unique to Thurstone, he was the first to demonstrate how it could be applied to pair-comparison data. If the perceptual representation of each individual stimulus is normal, then the distribution of the difference between two stimuli will also be normal and the probability of a correct judgment can be calculated based on the mean and variance of this distribution. Thurstone's model of comparative judgment has been shown to provide a reasonable fit to data when the stimuli are ordered on a unidimensional psychophysical continuum (e.g., judgments of weight). If the mean values for the normal distributions of stimulus perception (in Thurstone's model) are represented on a log scale (i.e., the mean perceptual magnitude is proportional to the logarithm of the physical magnitude), then this model will produce the Weber fraction. Our η model takes a similar approach in that we assume that there is variability in the quality of evidence extracted by the perceptual system and this evidence varies across trials according to a normal distribution.

7.3. Single stimulus tasks

If, as in the σ and η models, internal or external noise is related to overall stimulus strength, there is no reason to restrict this relationship to tasks with pairs of stimuli. However, previous applications of the diffusion model to single-stimulus tasks have generally not allowed either σ or η to vary across conditions and have produced adequate fits (although in memory tasks, across-trial variability in drift may increase with memory strength: Starns & Ratcliff, 2014; Starns et al., 2012). To investigate the relationship between these variability parameters and stimulus intensity, we re-fit data from a single stimulus brightness discrimination task (Ratcliff, 2014, Experiment 6) that had a wide range of brightness values (2.5–97.5 percent white pixels). In this task, subjects were presented with a single patch of black and white pixels and asked to judge whether there were more or less than 50% white pixels. To fit these data, for both the η and σ models we allowed drift rates to freely vary across brightness levels because, for this task and range of stimulus values, Ratcliff (2014) demonstrated that there is not a linear relationship between stimulus intensity and drift rates except in the middle of the range. For the σ model, we constrained within-trial variability in the accumulation process to be a linear function of the stimulus input as in Eq. (8) (where the S term is the proportion of white pixels and the s_1 parameter provides the linear mapping from stimulus value to within-trial variability). The intercept was fixed to 0.1 for scaling purposes.

$$\sigma_i = S_i s_1 + 0.1 \tag{8}$$

For the η model, we constrained across-trial variability in drift to be a linear function of the stimulus input as in Eq. (9) (where the S term is the proportion of white pixels and the e_i parameters provides the linear mapping from the stimulus value to within-trial variability).

$$\eta_i = S_i e_1 + e_2 \tag{9}$$

For comparison, we also fit a version of the η model where across-trial variability in drift was calculated in a manner more similar to the 2AFC version of the model, as shown in Eq. (10). In this version of the η model, across-trial variability in drift is a function of both the stimulus value and the criterion value (0.5) to which each single stimulus is compared.

$$\eta_i = e_1 \sqrt{S_i^2 + 0.5^2} + e_2 \tag{10}$$

Best-fitting parameter values from each model are shown in Table 7 (drift rates are not included to save space). For the σ model and the first version of the η model, while the parameters that control the mapping of stimulus intensity into variability (s_1 and e_1) do not entirely go to zero in these fits, the best-fitting values are quite small relative to the constant terms for these parameters (0.1 and e_2). For the σ model, values for σ across brightness conditions range from 0.100 to 0.108. For the first version of the η model, values for η across brightness conditions range from 0.062 to 0.069 (note that in the σ model the fixed value for η is slightly larger than this range at 0.072). For both η and σ, these ranges of values are quite small, less than an 11% change. For the second version of the η model, where the criterion value is included in the η formula, the parameter that controls the mapping of stimulus intensity into variability (e_1) was larger while the constant term (e_2) was smaller compared to the version of the model where η was a function of the stimulus strength alone. Including the criterion value in the equation thus produces results more similar to the 2AFC task in that there are slight changes in η across brightness levels and the constant term in the formula for η essentially goes to zero. Overall, these results indicate that these variability parameters do not seem to be related to stimulus intensity in a single-stimulus version of the brightness discrimination task. This may be because this task involves a comparison between a stimulus and some criterion value such

Table 7
Select parameter values from σ and η models for fit to single-stimulus task. a is the boundary separation, T_{er} is the mean non-decision time, s_z is the range of the across-trial variability of the starting point, p_z is the proportion of uniformly distributed contaminant responses, s_t is the range of the across-trial variability in non-decision time, z is the starting point of the evidence accumulation process, s_1 is the scaling parameter for within-trial variability in the σ model, η is the standard deviation for across-trial variability in drift in the σ model, and e_1 is the scaling parameter and e_2 the constant intercept for across-trial variability in drift rate in the η model.

σ model	a	T_{er}	s_z	p_z	s_t	z	s_1	η
	0.114	0.357	0.088	0.001	0.153	0.059	0.008	0.072
η model 1	a	T_{er}	s_z	p_z	s_t	z	e_1	e_2
	0.109	0.357	0.085	0.002	0.153	0.056	0.007	0.062
η model 2	a	T_{er}	s_z	p_z	s_t	z	e_1	e_2
	0.109	0.357	0.086	0.001	0.153	0.056	0.096	0.0001

that there may be additional pre-decision processing involved in this task (compared to the 2AFC tasks) before the decision-relevant information is fed into the decision process. This difference between the single-stimulus and two-alternative versions of the task is similar to the differences observed by Ratcliff and McKoon (in press) between several single-stimulus numerosity tasks and two-alternative tasks. (These models assume that variability is related to perceptual stimulus strength and not say the difference between the stimulus and the bright/dark criterion. Even if the model made variability a function of the difference, the changes in η and σ would be small.)

7.4. Theoretical considerations

Both the σ and η models were able to fit both the overall magnitude and stimulus difference effects and fit the observed data approximately equally well for our two stimulus tasks. Therefore on this basis, the two models cannot be discriminated. However, there are some arguments that the η model should be preferred over the σ model (that the authors are not in complete agreement about). If within-trial variability varies with stimulus strength, then it should do so for both single stimulus and two-stimulus tasks. First, the results above show that σ and η in their corresponding models vary to a significant degree only for the two stimulus tasks. If we believe the same principles apply in the numerosity domain also, then the σ model should apply in that domain also. In this domain, the σ model can fit the data when two arrays of dots are intermingled, but cannot fit data from the task in which two arrays are presented side by side. In the side by side task, both σ and η are approximately constant and drift rate varies as the log of numerosity and in a single stimulus task (is the number of dots greater or less than 25 say), again, both η and σ are approximately constant in their respective models. Thus whether within-trial noise in the decision process varies with stimulus strength is dependent on the experimental arrangement, not the fundamental properties of numerosity per se (i.e., not based on the numerosity magnitude).

The σ model also requires that σ is scaled to the experimental stimulus strength (drift rate also has to be scaled to stimulus strength) because in order to fit data from a task with weak stimuli and a task with strong stimuli, similar values of σ are needed for other parameters to be relatively invariant. In contrast, for the η model, drift rate has to be scaled to stimulus strength and that means that η is scaled by the same operation.

There is some debate over the presence of either of these parameters in models of decision-making (though one or the other is needed). The linear ballistic accumulator model, a popular competitor of the diffusion model, models the accumulation of evidence as a deterministic process with no noise in the accumulation process (Brown & Heathcote, 2008). Applications of the diffusion model in neurophysiological studies have not always included across-trial variability in drift rate (Ditterich, 2006a, 2006b; Palmer, Huk, & Shadlen, 2005; Shadlen & Kiani, 2013). However, this parameter is necessary to enable the model to produce slow error responses (Ratcliff & Rouder, 1998), even when the decision thresholds are allowed to decrease over time (Voskuilen, Ratcliff, & Smith, 2016).

In a more direct test of the need for across-trial variability in drift rate, Ratcliff, Voskuilen, and McKoon (2018) used a double-pass procedure to demonstrate that there is systematic variability across trials in the quality of evidence being used to make a decision. In double pass tasks, subjects were presented with identical stimuli on pairs of trials separated from each other by around 100 intervening trials. Responses to these pairs of stimuli can then be examined to see whether subjects make the same responses to repeated stimuli more often than would be predicted by chance alone. In terms of the diffusion model, if there is no across-trial variability in parameters then the variability in observed responses can only be the result of within-trial variability in the accumulation process. Because this source of variability is independent across presentations, if within-trial variability in the accumulation process is the only source of behavioral variability, then responses to pairs of stimuli in the double-pass procedure should not be in agreement more often than would be predicted by chance alone. However, if there is also across-trial variability in drift rate and at least some proportion of this variability is systematic (i.e., consistent across repetitions of the same stimuli) then responses to pairs of stimuli may be in agreement more often than predicted by chance. For five experiments with perceptual stimuli, Ratcliff et al. (2018) demonstrated that there is agreement between these separated repetitions that is greater than chance. These results support both the need to include across-trial variability in drift rate when applying this model and also that some amount of across-trial variability in drift may be related to stimulus properties (as in the η model).

The results presented here also parallel results from Ratcliff and McKoon (in press) described earlier that modeled processing in numerosity tasks with two arrays of stimuli. They developed two models for drift rates embedded in the diffusion decision model, models similar to those presented here. One had drift rates a linear function of the difference in numerosity between the two arrays and with across-trial SD in drift rate the square of the sum of squares of the two numerosities (both multiplied by coefficients) as in Eqs. (1) and (2). The other model assumed constant across-trial SD in drift rate and with drift rate the difference in the logs of the two numerosities. These are both models of representation in the numerical cognition literature. When the two arrays of dots were intermingled in a single area, the linear model fit the data better (the patterns were similar to those in Fig. 3). But when the two arrays were presented side by side, the log model fit the data best and plots of RT against accuracy fell on a single function unlike those in Fig. 3.

Ratcliff and McKoon argued that when separate representations of the two stimulus classes can be produced/estimated, then magnitude and difference effects affect drift rate and performance in the same way and interchangeably (i.e., changes in stimulus magnitude or in stimulus difference that produce the same change in accuracy will have the same effect on RT). Thus, drift rates represent the difference in magnitudes of the two separate stimuli (on a log scale in this case). In contrast, in intermingled arrays of dots, only local stimulus differences can be computed and these are accumulated over time. Thus, the linear model with increasing SD in drift rate with increasing magnitude fits data from this kind of task. For Experiments 1–5 presented here, although the stimuli are separated, the argument would be that separate representations cannot be produced for the two stimuli (unlike judgments of numerosity for separate static arrays, cf. no absolute identification) and differences are accumulated as for the numerosity task with

intermingled stimulus classes. Thus, the numerosity findings demonstrated that evidence representations may be task-dependent such that across-trial variability in drift may be related to stimulus intensity in some tasks but not others.

All these findings provide some theoretical support for the η diffusion model by demonstrating both support for a link between stimulus properties and across-trial variability in evidence entering the decision process and support for the hypothesis that this link is task-dependent. In contrast, the σ model is consistent with the assumption that the diffusion model represents a spike-counting process with stimulus intensity represented by neural firing rates (e.g., Simen et al., 2016).

8. Conclusions

Across five experiments and multiple conditions within each experiment, we observed effects of both overall stimulus magnitude and stimulus difference effects. These patterns of results are difficult for standard diffusion models to account for given that such models are generally only sensitive to relative information (i.e., stimulus differences) and not overall stimulus magnitude. We demonstrate that two versions of the diffusion model can account for these effects by allowing either within-trial noise in the accumulation process or across-trial noise in the quality of evidence to increase as a function of the overall stimulus magnitude. Although it performed well for some of the Experiments, we were not able to account for the entire range of effects using an LCA model. While a simple linear relationship between stimulus strength and perceptual strength was sufficient for most of our fits, different stimuli or ranges of stimulus values may require more complex transformations.

Acknowledgments

This work was supported by the National Institutes of Health, United States [R01-AG041176].

References

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review, 113*(4), 700–765.
Bowles, N. L., & Glanzer, M. (1983). An analysis of interference in recognition memory. *Memory & Cognition, 11*(3), 307–315.
Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology, 57*(3), 153–178.
Cohen, A. L., Rotello, C. M., & Macmillan, N. A. (2008). Evaluating models of remember-know judgments: Complexity, mimicry, and discriminability. *Psychonomic Bulletin & Review, 15*, 906–926.
Ditterich, J. (2006a). Evidence for time-variant decision making. *European Journal of Neuroscience, 24*, 3628–3641.
Ditterich, J. (2006b). Stochastic models of decisions about motion direction: Behavior and physiology. *Neural Networks, 19*, 981–1012.
Fechner, G. T. (1860). *Elemente der Psychophysik*. Breitkopf und Härtel.
Geisler, W. S. (1989). Sequential ideal-observer analysis of visual discriminations. *Psychological Review, 96*(2), 267–314.
Heathcote, A. (1998). Neuromorphic models of response time. *Australian Journal of Psychology, 50*(3), 157–164.
Heathcote, A. (2003). Item recognition memory and the receiver operating characteristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 1210–1230.
Hirshman, E., & Hostetter, M. (2000). Using ROC curves to test models of recognition memory: The relation between presentation duration and slope. *Memory & Cognition, 28*, 161–166.
Hunt, L. T., Kolling, N., Soltani, A., Woolrich, M. W., Rushworth, M., & Behrens, T. (2012). Mechanisms underlying cortical activity during value-guided choice. *Nature Neuroscience, 15*(3), 470–476.
MacKay, D. M. (1963). Psychophysics of perceived intensity: A theoretical basis for Fechner's and Stevens' laws. *Science, 139*, 1213–1216.
Marshall, J. A. R., Bogacz, R., Dornhaus, A., Planqué, R., Kovacs, T., & Franks, N. R. (2009). On optimal decision-making in brains and social insect colonies. *Journal of the Royal Society, Interface/The Royal Society, 6*(40), 1065–1074.
Mickes, L., Wixted, J. T., & Wais, P. E. (2007). A direct test of the unequal-variance signal-detection model of recognition memory. *Psychonomic Bulletin & Review, 14*, 858–865.
Miletić, S., Turner, B. M., Forstmann, B. U., & van Maanen, L. (2017). Parameter recovery for the Leaky Competing Accumulator model. *Journal of Mathematical Psychology, 76*(A), 25–50.
Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *Computer Journal, 7*, 308–313.
Niwa, M., & Ditterich, J. (2008). Perceptual decisions between multiple directions of visual motion. *Journal of Neuroscience, 28*, 4435–4445.
Palmer, J., Huk, A. C., & Shadlen, M. N. (2005). The effect of stimulus strength on the speed and accuracy of a perceptual decision. *Journal of Vision, 5*, 376–404.
Pirrone, A., Azab, H., Hayden, B. Y., Stafford, T., & Marshall, J. A. R. (2017). Evidence for the speed-value trade-off: Human and monkey decision making is magnitude sensitive. *Decision*. Advance online publication. http://dx.doi.org/10.1037/dec0000075.
Polania, R., Krajbich, I., Grueschow, M., & Ruff, C. C. (2014). Neural oscillations and synchronization differentially support evidence accumulation in perceptual and value-based decision making. *Neuron, 82*, 709–720.
Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*, 59–108.
Ratcliff, R. (2013). Parameter variability and distributional assumptions in the diffusion model. *Psychological Review, 120*, 281–292.
Ratcliff, R. (2014). Measuring psychometric functions with the diffusion model. *Journal of Experimental Psychology: Human Perception and Performance, 40*, 870–888.
Ratcliff, R., & McKoon, G. (in press). Modeling numerosity representation using an integrated diffusion model. *Psychological Review*.
Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation, 20*, 873–922.
Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science, 9*, 347–356.
Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling modeling for two-choice reaction time. *Psychological Review, 111*, 333–367.
Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin and Review, 9*, 438–481.
Ratcliff, R., Voskuilen, C., & McKoon, G. (2018). Internal and external sources of variability in perceptual decision-making. *Psychological Review, 125*, 33–46.
Shadlen, M. N., & Kiani, R. (2013). Decision making as a window on cognition. *Neuron, 80*, 791–806.
Shepard, R. N. (1981). On the status of 'direct' and psychophysical scales: Psychophysical measurement. *Journal of Mathematical Psychology, 24*, 21–57.
Simen, P., Vlasov, K., & Papadakis, S. (2016). Scale (in)variance in a unified diffusion model of decision making and timing. *Psychological Review, 123*(2), 151–181.
Smith, P. L., & Ratcliff, R. (2009). An integrated theory of attention and decision making in visual signal detection. *Psychological Review, 116*, 283–317.
Starns, J. J., & Ratcliff, R. (2014). Validating the unequal-variance assumption in recognition memory using response time distributions instead of ROC functions: A diffusion model analysis. *Journal of Memory and Language, 70*, 36–52.
Starns, J. J., Ratcliff, R., & McKoon, G. (2012). Evaluating the unequal-variance and dual-process explanations of zROC slopes with response time data and the

diffusion model. *Cognitive Psychology, 64*(1–2), 1–34.
Stevens, S. S. (1957). On the psychophysical law. *Psychological Review, 64*(3), 153–181.
Teodorescu, A. R., Moran, R., & Usher, M. (2016). Absolutely relative or relatively absolute: Violations of value invariance in human decision making. *Psychonomic Bulletin & Review, 23*(1), 22–38.
Teodorescu, A. R., & Usher, M. (2013). Disentangling decision models: From independence to competition. *Psychological Review, 120*(1), 1–38.
Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*(4), 273–286.
Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review, 108*(3), 550–592.
Vickers, D. (1970). Evidence for an accumulator model of psychophysical discrimination. *Ergonomics, 13*, 37–58.
Voskuilen, C., Ratcliff, R., & Smith, P. L. (2016). Comparing fixed and collapsing bound versions of the diffusion model. *Journal of Mathematical Psychology, 73*, 59–79.