Cognition 152 (2016) 170–180

Contents lists available at ScienceDirect

Cognition

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

Original Articles

Strength and weight: The determinants of choice and confidence ☆

Peter D. Kvam [a,b,*], Timothy J. Pleskac [b]

[a] Michigan State University, United States
[b] Max Planck Institute for Human Development, Germany

ARTICLE INFO

Article history:
Received 13 October 2015
Revised 5 April 2016
Accepted 11 April 2016
Available online 16 April 2016

Keywords:
Evidence accumulation
Strength
Weight
Confidence
Decision-making
Diffusion model

ABSTRACT

Evidence for different hypotheses is often treated as a singular construct, but it can be dissociated into two parts: its strength, the proportion of pieces of information favoring one hypothesis; and its weight, the total number of pieces of information available. However, cognitive and neural models of evidence accumulation often make a *proportional representation* assumption, implying that people take these two factors into account equally when making their decisions and judgments. We examine this assumption by directly manipulating the number of samples and the proportion favoring either of two alternatives in dynamic decision making and judgment tasks. The results suggest that people tend to over-emphasize the strength of evidence relative to its weight in both an optional-stopping decision task and a probability judgment task. In a drift-diffusion model, this is reflected by drift rates that are determined foremost by strength with a smaller influence of weight. This result challenges the proportional representation assumption made by existing models of judgment and decision-making, and calls into question modeling evidence accumulation as a Bayesian belief updating process.

1. Introduction

Ordinarily, information or evidence is thought of as a singular construct, one which informs our beliefs about and guides our actions in choosing between hypotheses or alternatives. However, evidence can often be dissociated into two components. These are the extremeness or proportion of instances where it favors a particular hypothesis — the *strength* of the evidence — and the total amount or reliability of the data — the *weight* of the evidence. Consider a simple example that Griffin and Tversky (1992) used to illustrate the distinction:

> Imagine that you are spinning a coin, and recording how often it lands heads and tails. Unlike tossing, which (on average) yields an equal number of heads and tails, spinning a coin leads to a bias favoring one side or the other because of slight imperfections on the rim of the coin (an uneven distribution of mass). Now imagine that you know that this bias is 3/5. It tends to land on one side 3 out of 5 times. But, you do not know if this bias is in favor of heads or in favor of tails.

The evidence collected via spinning the coin informs a person's belief that the bias is in favor of heads or tails. The weight in this case is the number of spins (the sample size), and the strength of the evidence is the proportion of times the coin came up heads. Bayes' rule implies both the weight and the strength of the evidence should have equal importance in determining the confidence of the bias in the coin.

Critically, Griffin and Tversky (1992) found that when participants judged the probability of a heads bias relative to tails, these judgments were influenced more by changes in the strength than changes in the the weight of available evidence. This greater influence of strength over weight not only included judgments about chance (or aleatory) events, but also beliefs about epistemic events (i.e., whether a fact was true or not) as well.

The standard explanation for why strength carries more influence on judgments than weight resides in a dual-process framework (Kahneman, 2003; Sloman, 1996). That is, case-based information, such as sample proportion, is intuitively and rapidly assessed by a heuristic system (System 1), while accurate computation of likelihoods integrating class-based weight information requires the action of a second, more deliberative system (System 2) (Brenner, Griffin, & Koehler, 2012). This theory suggests that the gap arises, particularly at short time intervals, because System 1 has more time to operate and therefore processes more information than does System 2, resulting in an emphasis on strength relative to weight information in observed responses.

☆ This research was supported by a grant from the National Science Foundation (SBE-0955140), and the first author was additionally supported by a graduate fellowship from the National Science Foundation (DGE-1424871).

* Corresponding author at: Department of Psychology, Michigan State University, East Lansing, MI 48824, United States.
E-mail address: kvam.peter@gmail.com (P.D. Kvam).

http://dx.doi.org/10.1016/j.cognition.2016.04.008
0010-0277/

However, it is important to note that this dual-system view of the strength-weight gap is based mainly on static, high-level cognitive tasks where information is readily available and can be processed in any order. Two interesting variations, then, are simple perceptual decisions and cases where information arrives or is gathered dynamically. Finding a strength-weight gap in the perceptual decisions would suggest that these two dimensions of information are treated differently on a more fundamental level, perhaps due to differences in sensitivity to each factor (see for example Feigenson, Dehaene, & Spelke, 2004; Gallistel & Gelman, 2000; Longo & Lourenco, 2007). And evidence for a strength-weight gap in experience-based decisions, given the differences in risky choice behavior between experience- and description-based decisions (Hertwig, Barron, Weber, & Erev, 2004; Hertwig & Erev, 2009), would suggest that it is a robust phenomenon across different methods of information presentation. The experiments we outline below test each of these possibilities.

Curiously, most computational models of judgment and decision making tend *not* to make an explicit distinction between the strength and weight of evidence and thus implicitly assume equal emphasis on both dimensions. Sequential sampling models — arguably the most successful at predicting decisions, response times, and judgments — instead represent evidence as a tally or sum of pieces of information extracted from the stimulus itself or from some cognitive representation of the object or item in question (e.g., Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006; Busemeyer & Townsend, 1993; Pleskac & Busemeyer, 2010; Ratcliff, 1978; Ratcliff & McKoon, 2008; Smith & Van Zandt, 2000; Usher & McClelland, 2001). This assumption about the nature of evidence is a result of the *proportional representation* assumption that these models make. This assumption postulates that the decision maker represents and updates evidence for potential hypotheses in a way that (noisily) mirrors the actual characteristics of the stimulus. In many cases, the rate of evidence accumulation is even set directly from the features of the stimuli (e.g., Busemeyer & Townsend, 1993; Krajbich, Lu, Camerer, & Rangel, 2012; Link & Heath, 1975; Nosofsky & Palmeri, 1997; Palmer, Huk, & Shadlen, 2005; Ratcliff, 2014).

The proportional representation assumption has its roots in the sequential probability ratio test (SPRT) framework, on which these models are based. According to the SPRT framework, the evidence state at any give time approximates the log likelihood of the possible hypotheses given the accumulated information, usually to within some scaling factor (Bogacz et al., 2006; Edwards, 1965; Wald & Wolfowitz, 1949). The evidence state is then optimally updated with each new piece of information according to Bayes rule. Probability judgments can then be made accurately from this representation, and decisions are made when the evidence exceed a threshold magnitude (e.g. ± 5). The height of the threshold may correspond to a desired level of confidence required to make a decision.

Reliance on the SPRT framework and the proportional representation assumption extends to neural models as well (Kira, Yang, & Shadlen, 2015). Recent neural models of decision making assume evidence is represented in the brain in terms of (or approximates) the relative log odds of the hypotheses (Beck et al., 2008; Gold & Shadlen, 2001; Kiani & Shadlen, 2009; Meyniel, Sigman, & Mainen, 2015) and that the odds are updated in a manner consistent with Bayes rule (Kepecs, Uchida, Zariwala, & Mainen, 2008; Knill & Pouget, 2004; Ma, Beck, Latham, & Pouget, 2006). While neural representations of evidence almost certainly reflect a person's beliefs about the hypotheses, findings indicating unequal credence given to strength and weight in judgments or decisions would suggest that the evidence used to make these responses cannot be updated in an optimal Bayesian way.

Most current theories suggest that both choices and confidence judgments use the same underlying evidence representations (Kiani & Shadlen, 2009; Merkle & Van Zandt, 2006; Moran, Teodorescu, & Usher, 2015; Pleskac & Busemeyer, 2010; Ratcliff & Starns, 2013; Van den Berg et al., 2016; Vickers, 1979; Yu, Pleskac, & Zeigenfuse, 2015). This means that an imbalance of strength and weight contributions to evidence representations should yield differences in both confidence judgments as well as choice proportions and response times. We therefore test both response types in our experiments. In the next section, we establish model predictions for both choice and judgment tasks by examining theories of how strength and weight are combined to generate the evidence underlying choice and confidence.

1.1. Definitions and proofs

Evidence in sequential sampling models is a function of both the weight and strength. However, the precise function depends on if the sequential sampling model uses a relative stopping rule or if it uses an absolute stopping rule (Ratcliff & Smith, 2004). A relative stopping rule implies that evidence from the stimulus in favor of one response alternative is evidence against the other alternative, with evidence accumulation halting to make a decision when the balance between hypotheses reaches a threshold. Absolute stopping rules, in comparison, imply that increasing evidence for one alternative does not change the evidence for another alternative, and a decision is triggered when the evidence for a single hypothesis reaches a threshold. In what follows, we show that in sequential sampling models using relative stopping rules (i.e., random walk and their continuous time variants, drift diffusion models) evidence is equally determined by strength and weight. Appendix A shows that this same result is true for models with absolute stopping rules.

In order to examine the predictions of sequential sampling models as they relate to strength and weight, we first establish operational definitions of each of these constructs. For our purposes, we define the weight of incoming information to be the number of samples that a person receives. We define strength as a linear transformation of the proportion of samples that favor one alternative. In this case, we define it relative to alternative A (e.g. coin is 2/3 heads), so that positive values for strength indicate that more samples favor A, and negative values of strength indicate that more samples favor B. More formally, setting $A(t)$ and $B(t)$ as the total number of observations at time t in favor of each alternative, we define weight and strength as follows:

$$\begin{aligned} &\textbf{Weight}: \quad w(t) = A(t) + B(t) \\ &\textbf{Strength}: \quad s(t) = \frac{A(t) - B(t)}{A(t) + B(t)} \end{aligned} \tag{1}$$

In a sequential sampling model using a relative stopping rule, the evidence state $P(t)$ is represented as the number of samples favoring one option minus the number favoring the other, $P(t) = A(t) - B(t)$. Substituting our definitions from Eq. (1), this position can also be represented in terms of strength and weight.

$$P(t) = A(t) - B(t) = w(t) \cdot s(t) \tag{2}$$

One consequence of defining $P(t)$ as the difference in number of samples is that it can be understood as approximating the log likelihood of the samples given hypothesis A relative to hypothesis B. According to Bayes' rule, the true posterior log odds of hypothesis A relative to B can be computed by combining the likelihood with prior probabilities of the two hypotheses.

$$\frac{Pr(A|D)}{Pr(B|D)} = \frac{Pr(D|A)}{Pr(D|B)} \cdot \frac{Pr(A)}{Pr(B)} \tag{3}$$

Assuming both hypotheses are initially equally likely, $\frac{Pr(A)}{Pr(B)} = 1$, we can represent the log transformed posterior odds on the lefthand side of Eq. (3) as a straightforward transformation of $P(t)$ (see Edwards, 1965; Wald & Wolfowitz, 1949):

$$\ln\left(\frac{Pr(A|D)}{Pr(B|D)}\right) = P(t)\cdot\ln(d') = s(t)\cdot w(t)\cdot\ln(d') \quad (4)$$

The term d' is the discriminability of the alternatives as found in signal detection theory (Green & Swets, 1966) – for example, if a coin is either 70% heads [A] or 30% heads [B], d' would be $\frac{7}{3}$). Thus, probability judgments about the relative credibility of the two hypotheses, insofar as they are scaled from the internal evidence state $P(t)$, take strength and weight equally into account. This is where the models appear to come into conflict with the data from Griffin and Tversky (1992).

However, note that so far we can only predict confidence judgments for a fixed set of data. To extend the model predictions to dynamic choice and confidence data, we must specify two additional properties. First of these is the rate of accumulation of the evidence for hypothesis A, known as the *drift rate* (μ). This corresponds to the rate of change of $P(t)$, which can be computed in terms of strength and weight by taking the derivative of Eq. (2) with respect to time.

$$P'(t) = w'(t)\cdot s(t) + w(t)\cdot s'(t) \quad (5)$$

We can compute the expected value of the drift more simply by noting that strength does not change systematically with time (i.e. $E[s'(t)] = 0$) as long as samples are generated from a consistent source.

$$E(\mu) = E[P'(t)] = E[w'(t)\cdot s(t)] \quad (6)$$

As Eq. (6) shows, in dynamic models of judgment and decision making the drift rate or change in evidence per unit of time should equally emphasize the strength of the evidence at each time point and the rate of change in the weight of the evidence. The experiments we describe next test this prediction for dynamic choices and confidence judgments by directly modeling the effects of manipulations of strength and weight on the drift rate in both choice and confidence responses.

However, before we can model dynamic choices we need to include a second extension, the *decision threshold*. The decision threshold is used to determine when a choice is made and what is chosen, such that a decision is triggered when the evidence $P(t)$ exceeds the decision threshold θ. More specifically, a person will choose alternative A once $P(t) \geqslant \theta$ and choose alternative B once $P(t) \leqslant -\theta$. The decision threshold θ is important because it allows one to separate evidence accumulation rates from the total amount of evidence collected. Note separate bounds for alternative A and B can be specified if we wish to introduce prior bias into the accumulation process; however, this is unnecessary given our experimental design.[1]

[1] Note the generality of our definition of the weight $w(t)$ and strength $s(t)$ of evidence implies the prediction of equal emphasis on strength and weight applies to a wide range of sequential sampling models including those that (a) treat evidence as some function of the likelihood of the information in respect to the different response alternatives (e.g., Edwards, 1965; Laming, 1968; Van den Berg et al., 2016); (b) based on a comparison between the sampled information and a mental standard (e.g., Link & Heath, 1975); (c) evidence scaled proportionally from the observed stimulus (Palmer et al., 2005; Ratcliff, 2014) or (d) a measure of strength based on a match between a memory probe and memory traces stored in long-term memory (Ratcliff, 1978).

1.2. Predictions and studies

All of the models we have covered predict that drift rates in both confidence and choice tasks should be coequally determined by strength and weight. This includes basic random walk models (Edwards, 1965; Laming, 1968; Link & Heath, 1975; Kira et al., 2015; Stone, 1960), the drift-diffusion model and its extensions (Busemeyer & Townsend, 1993; Pleskac & Busemeyer, 2010; Ratcliff, 1978; Ratcliff & McKoon, 2008; Ratcliff & Starns, 2009), neural models of evidence accumulation (Beck et al., 2008; Gold & Shadlen, 2001; Kiani & Shadlen, 2009; Kira et al., 2015; Knill & Pouget, 2004; Ma et al., 2006; Palmer et al., 2005) and even accumulator-based models with absolute rather than relative stopping rules (Smith & Van Zandt, 2000; Usher & McClelland, 2001) (see Appendix A).

To test these predictions, we created a dynamic version of Griffin and Tversky (1992)'s coin example. During the task, pink or green dots appeared one-by-one on a computer screen. For a given trial, the dots were drawn from a population of dots that either had more green or more pink dots. Participants either had to identify if the dots were being drawn from the pool with more green or pink dots (choice task), or rate their confidence (via a probability judgment) that the dots were being drawn from a pool with more pink dots (confidence task). In both the choice and confidence conditions, we manipulated the strength and weight of this information by changing the relative frequency and rate of arrival of the dots, respectively. To equate confidence and choice tasks, we yoked the choice and confidence conditions. In particular, the dot sequence and response times from the choice tasks were used to determine the sample of dots shown and the presentation time of the confidence condition.

We used the drift diffusion model as a measurement model, examining the effect of strength and weight manipulations on estimated drift rates in choice and confidence tasks (Ratcliff, 2014; Voss, Rothermund, & Voss, 2004). In a perfect Bayesian belief updating case, drift should always load entirely onto the interaction of strength and weight. However, spurious linear components may be introduced if errors are not homogeneous over the range of strength and weight levels, so we allowed for separate linear components of strength and weight. If the proportional representation assumption holds, then the effects of manipulations of strength and weight should be the same and any linear components should be balanced between strength and weight, as they result from error related to their joint function. Alternatively, it may be that even during optional stopping and probability judgment tasks with dynamic stimuli that, just as Griffin and Tversky (1992) found with static fixed sample inferential tasks, participants over-emphasize the strength of the evidence relative to its weight. As we reviewed earlier, such a result would challenge existing dynamic decision models that assume evidence accumulation is veridically a Bayesian belief updating process.

2. Methods

2.1. Participants

A total of 29 Michigan State University undergraduate students participated in the experiment for class credit. Participants were 69% female (31% male) and primarily 18–26 years old. Each participant completed each of the 3 tasks – choice, confidence, and numerical estimates – which combined took approximately 1 h to complete.

2.2. Materials

The task was programmed in MATLAB using Psychtoolbox 3 (Brainard, 1997; Kleiner et al., 2007). Participants were seated

individually in sound-dampening booths for the entirety of the experiment following the initial briefing. All responses were recorded using the mouse.

2.2.1. Stimuli

The stimuli were generated in MATLAB and Psychboolbox and included pink or green circles, each occupying approximately 0.4 visual degrees, arranged within an aperture with a visual angle of approximately 10 degrees on a black background. During the task, dots appeared one by one on the screen at a rate depending on the weight manipulation, which included rates of 3, 5, 7, 11, or 15 dots per 2 s. The proportion of dots that were pink or green were generated from 4 levels of strength which covered sample proportions of 51%, 65%, 79%, and 93% (strength levels of 0.02, 0.30, 0.58, and 0.86, respectively). In the choice condition, dots continued to accumulate until a participant made their response. In all other conditions, the dots accumulated for a set period of time before disappearing from the screen.

2.3. Choice task

In the choice task, participants saw the dot stimuli appear one by one and accumulate on the screen (see Fig. 1, top row). Participants were told that these dots were being pulled randomly from a larger pool of dots that was either 2/3 green and 1/3 pink or 2/3 pink and 1/3 green (i.e., discriminability was held constant across the experiment at $d' = 2$). Whenever participants were ready to answer, they clicked the left or right mouse button to indicate whether they believed the dots were coming from the 2/3 green or the 2/3 pink pool, respectively. We recorded the precise number and order of dots on each trial as well as the timing and accuracy with which participants responded.

2.3.1. Analysis

The response time and accuracy data from the choice task were analyzed using the drift-diffusion model as a measurement model (Ratcliff, 2014; Voss et al., 2004). This model is frequently used to describe the evidence accumulation process which is posited to underlie both decision-making and judgments (Busemeyer & Townsend, 1993; Pleskac & Busemeyer, 2010; Ratcliff, 1978; Ratcliff & McKoon, 2008), and it has a number of advantages over standard linear statistical models. Critically, it controls for important characteristics of the evidence accumulation process such as the time it takes to execute a response (non-decision time), the within-trial variability in evidence, and the skewed distributions of response times, all of which can interfere with fits based on simple linear predictions.

The diffusion model was implemented using a hierarchical Bayesian implementation of the Wiener diffusion model provided by Wabersich and Vandekerckhove (2014), using JAGS and the matjags interface (see also Vandekerckhove, Tuerlinckx, & Lee, 2011). This model includes a drift rate parameter μ, which corresponds to the rate of change of evidence described in Eq. (6). Drift was set as a linear function of strength, weight, and their interaction. The coefficients on these factors (designated by δs) were set hierarchically by participant. The estimates of these coefficients allowed us to directly test the hypothesis that the rate of evidence accumulation is equally determined by the strength and weight of incoming evidence.

Alongside drift, the model also included a diffusion parameter σ^2 that describes the variance in evidence, allowing us to examine how the variability in accumulated evidence changes as a function of strength and weight manipulations as well. The intercept of this parameter was fixed in order to set the scale of the diffusion process, but it was also permitted to vary as a function of strenght, weight, and their interaction, with coefficients (designated by ϵ) again set hierarchically by participant. Note that this parameter is often fixed to 1 across conditions, but this unnecessarily restricts the model and can lead to variation being absorbed into other parameter estimates, obscuring the true locus of variability (Donkin, Brown, & Heathcote, 2009).

The choice model included two additional free parameters. This included the threshold θ, which was also set as a hierarchical linear function of strength and weight with a fixed intercept. Finally, non-decision time, a parameter representing the length of response time during which participants were not accumulating evidence, was set hierarchically by participant.

The JAGS model code is provided in Appendix B.

2.4. Confidence task

In the confidence task, participants saw the same dot stimuli as in the choice task. Because we were interested in whether the type of response (choice versus confidence) affected participants' use of the same evidence, the dots in the confidence task were exactly matched to those that they saw in the choice task. This was done by pulling the order of dots and response times directly from participants' prior responses on the choice task. For example, if participants responded in 2.4 s after seeing 'green, green, pink, green,' [GGPG] in the choice task, they would again see this same pattern of dots or its precise inverse [PPGP] over 2.4 s in the confidence task. The order was randomly determined. This yoking procedure enabled us to measure response times and confidence ratings for the same stimuli and examine how they related to one another as well as to the stimulus information.

After all dots had been presented, they were removed from the screen and a confidence scale appeared (see Fig. 1, middle row). Participants then rated the probability that the dots had come from the 2/3 pink pool, from 0 (completely sure of 2/3 green) to 100 (completely sure of 2/3 pink), by clicking on the circular scale.

2.4.1. Analysis

To model confidence judgments, we assumed that the probability judgment was a logistic (inverse logit) function of the accumulated evidence at time t in the drift diffusion model. This is a simplifying assumption that corresponds to current sequential sampling models of confidence judgments Pleskac and Busemeyer (2010). Because of this logit space assumption and because responses were externally cued, the model used to predict confidence estimates is simpler than that used for choice. Instead, the only free parameters in this model are drift (μ) and diffusion (σ^2), which give the mean and variance of a normal distribution of log odds judgments, respectively. As in the choice task, these parameters were set as linear functions of strength, weight, and their interaction, whose coefficients were in turn set hierarchically by participant.

Unlike in the choice trials, however, the strength and weight values were pulled from the actual stimuli. Because we were predicting log odds judgments from after the stimuli had been fully presented, we had to use the *total* weight of information that participants saw rather than the weight per unit time. Correspondingly, the strength coefficient was also set based on the stimulus characteristics rather than the seed value from the choice trial on which it was based. This has the same effect as computing the drift rate and noise and then multiplying them by the length of time for which the stimulus was displayed, so the interpretation of these parameter estimates is essentially the same as in the choice task.

The JAGS code and fitting details for this model are also provided in Appendix B.

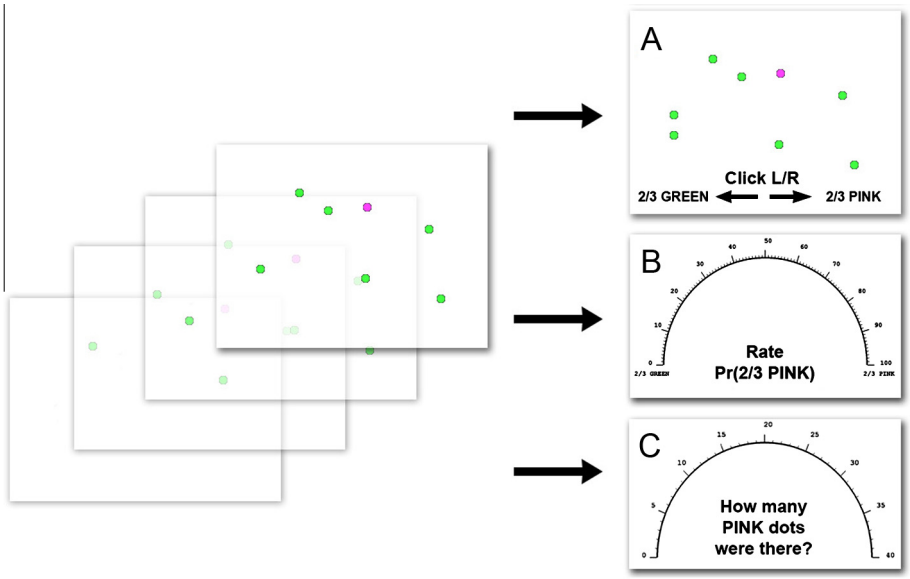

Fig. 1. Diagram of the choice (A), confidence (B), and number estimation (C) tasks that participants would perform.

2.5. Numerical estimation task

To gauge how well participants gathered and stored the information presented to them during the experiment, each participant completed an estimation task between blocks of choice and confidence trials. As in the other tasks, participants saw pink and green dots accumulate one by one on the screen. After 2 s, all of the dots were removed and participants were asked to estimate on a circular scale (see Fig. 1, top row) how many pink, green, or how many total dots had appeared on the screen during that time. For each trial, a random draw determined whether a participant was asked to report the number of pink, green, or total dots.

2.6. Procedure

Participants were briefed initially on all 3 tasks that they would see, including choice, confidence, and estimation tasks. Once seated for the experiment, they were shown examples of 2/3 green as well as 2/3 pink pools by watching 200 dots with the matching frequencies appear on the screen. Following this, participants completed training on the choice task and then 8–10 blocks of 20 trials of the choice task.[2] After each block, they also completed 4 trials of the estimation task. Once participants completed the choice section of the study, they were trained on the confidence task and then completed 8–10 blocks of 20 confidence trials. Although the presentations times and sequence of dots in the confidence task were yoked to the prior choice trials, the order of confidence trials was randomized relative to the choice task. After each block of the confidence task, participants again completed 4 trials of the estimation task.

3. Results

For all parameter estimates reported in this section, we present the mean estimate as well as the 95% highest density interval [HDI], which spans the 95% most credible values based on the posterior estimates of the parameter values. For the model predictions for accuracy, response times, and confidence, we calculated the predicted accuracy, mean response time, and mean confidence for each sample of parameters of the MCMC chain. Figs. 2 and 3 plot the mean of these mean model predictions along with the 95% highest density interval of the mean predictions.

We used Bayesian estimation techniques to estimate the models for choice, confidence and numerical estimation (Kruschke, 2010; Lee & Wagenmakers, 2013). Unless otherwise specified, all posterior parameter estimates are calculated using a likelihood with a diffuse prior consistent with those used by previous authors Kruschke (2010) and Wabersich and Vandekerckhove (2014) so as to let the data have maximal influence on the posterior estimates. Choice and confidence model parameters were estimated using 8 parallel chains. Each chain was comprised of 1000 burn-in steps (unrecorded samples to allow the chain to reach the reasonable parameter space) and 5000 samples. Preliminary analyses confirmed that all chains converged.

3.1. Choice task

Recall that the choice model used drift (μ), diffusion (σ^2), threshold (θ), and non-decision time (ndt) parameters. Because response alternatives were symmetric and because there was no apparent tendency to favor one color or the other, we fixed the bias in the model to 0.5 (unbiased). The posterior model fits are displayed in Fig. 2, which shows that the model provided a reasonable fit to the accuracy and response time data.

The estimates of the coefficients in the model are shown in Table 1. The coefficients for the drift rate are designated by δ (e.g., $\delta_{s\times w}$ corresponds to the coefficient on the interaction between strength and weight), while coefficients for noise are designated by ϵ. There are several things to note from these results: first, strength ($M(\delta_s) = 0.85$, 95% HDI = $[0.77, 0.92]$ has a much greater impact on drift than weight does ($M(\delta_w) = 0.48$, 95% HDI = $[0.42, 0.55]$). On average, manipulations of strength had approximately $1.8\times$ the effect of manipulations of weight (ratio $M(\delta_s : \delta_w) = 1.77$, 95% HDI = $[1.56, 1.98]$). This violates the predicted relationship given in Eq. (6), suggesting that participants are indeed using strength information more than weight when making their decisions.

Second, the coefficients for the threshold reveal that both strength ($M(\zeta_s) = 0.62$, 95% HDI = $[0.37, 0.85]$) and weight

[2] Early study participants completed only 8 blocks, but this was taking an insufficient amount of time to complete, so we increased the number of trials to 10 blocks of both choice and confidence trials for later participants.

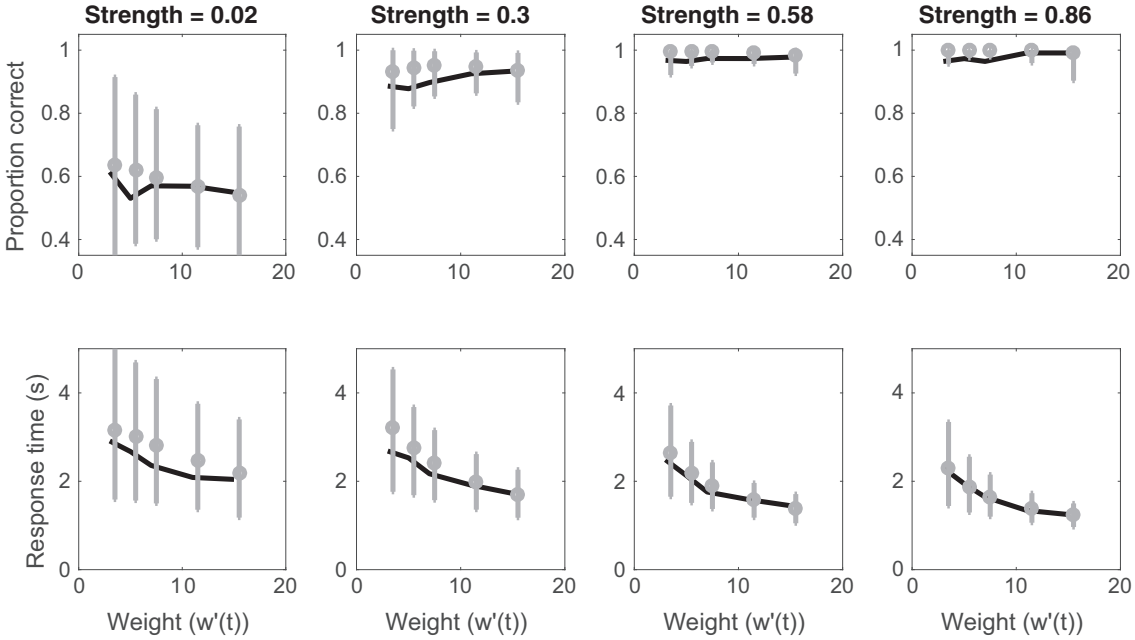

Fig. 2. Mean accuracy and response times in the data (black) and posterior mean of diffusion model predictions with 95% HDI (gray) for accuracy and response times across strength and weight manipulations.

($M(\zeta_w) = 0.11$, 95% HDI = $[0.00, 0.25]$) are positive predictors of threshold. These manipulations were not known to a participant before stimulus onset; therefore, these indicate on-line threshold changes, speaking against the static threshold assumption made by most sequential sampling models.[3]

Finally, noise in this model increases minimally with strength ($M(\epsilon_s) = 0.11$, 95% HDI = $[0.02, 0.20]$) and the strength-weight interaction ($M(\epsilon_{s\times w}) = 0.08$, 95% HDI = $[0.04, 0.12]$), but increases quite substantially with weight ($M(\epsilon_w) = 0.23$, 95% HDI = $[0.19, 0.28]$). This suggests that a higher number of evidence samples results in greater variability in evidence representation. This result is consistent with the finding that mental representations of larger numerical magnitudes have lower precision (Feigenson et al., 2004; Gallistel & Gelman, 2000).

3.2. Judgment task

Recall that the confidence model used a drift rate (μ) and a diffusion parameter (σ^2), which were set as a hierarchical linear function of strength and weight manipulations. The model predictions and data for the log odds of participants' judgments across strength and weight conditions are shown in Fig. 3. Also shown are the true Bayesian posterior log odds of the hypotheses, shown as the dotted black line. Note that participants' estimates are generally under-confident, especially when strength was high and weight was low (similar to the results of Griffin & Tversky, 1992). They also tend to be under-confident relative to their own accuracy, which was close to 100% in most conditions.

The group-level mean estimates for the parameters in the model are displayed in Table 2. As in the choice task, these estimates indicated that drift was more heavily influenced by strength ($M(\delta_s)$ = 0.37, 95% HDI = $[0.27, 0.48]$) than by weight ($M(\delta_w)$ = 0.12, 95% HDI = $[0.05, 0.20]$) or their interaction ($M(\delta_{s\times w})$ = 0.11, 95% HDI = $[0.07, 0.15]$). On average, manipulations of strength had approximately 3× the effect on confidence relative to manipulations of weight ($M(\delta_s : \delta_w)$ = 3.28, 95% HDI = $[1.45, 6.23]$). Note the HDIs of this proportion overlap substantially between the choice and confidence conditions, indicating that the relative impacts of strength and weight are credibly the same between tasks.

As in the choice task, diffusion changed substantially with weight ($M(\epsilon_w)$ = 0.54, 95% HDI = $[0.10, 0.98]$), and the strength-weight interaction ($M(\epsilon_{s\times w})$ = 0.78, 95% HDI = $[0.22, 1.32]$), but not with strength alone ($M(\epsilon_s)$ = 0.24, 95% HDI = $[-0.00, 0.47]$). The increase in dispersion of log odds estimates with strength and weight can be seen clearly in Fig. 3. Note that these effects are likely to be somewhat inflated because motor errors between ratings high on the scale result in much larger variance in log odds than errors between ratings low on the scale. For example, responding at 95–99 when one means to respond at 97, when the probabilities are converted to log odds, yields much larger deviations in log odds space compared to responding on 65–69 when one means to respond at 67. However, this does not fundamentally change the significance of the observation that higher weight abnormally increases evidence variability.[4]

3.3. Numerical estimation task

In the estimation task, we examined how well people we able to store and recall the number of pink, green, and total dots that actually appeared on the screen. The precision of participants'

[3] It is worth noting that fixing the threshold parameter does not substantially change the conclusions we draw based on estimates of other parameters in the model (i.e., they don't just result from over-fitting). In such a model, strength ($M(\delta_s) = 0.80$, 95% HDI = $[0.76, 0.85]$) still affects drift moreso than weight ($M(\delta_w) = 0.49$, 95% HDI = $[0.42, 0.57]$) and their interaction ($M(\delta_{sw}) = 0.35$, 95% HDI = $[0.29, 0.41]$).

[4] Note that the study reported here is not counterbalanced because we wanted to match the choice and confidence conditions. However, in a previous study where the confidence trial duration was fixed at 2 s long and counterbalanced with choice, we found very similar results (with some differences reflecting the caveat that strength and weight manipulations were range restricted relative to the yoked confidence condition): the estimates for the confidence task suggested that drift had a credibly zero intercept ($M(\delta_0) = 0.00$, 95% HDI = $[-0.28, 0.28]$) and increased with strength ($M(\delta_s) = 0.27$, 95% HDI = $[0.22, 0.32]$) moreso than weight ($M(\delta_w) = 0.09$, 95% HDI = $[0.01, 0.18]$) or their interaction ($M(\delta_{s\times w}) = 0.16$, 95% HDI = $[0.12, 0.20]$). Diffusion increased with weight ($M(\epsilon_w) = 0.24$, 95% HDI = $[0.11, 0.35]$) but credibly didn't increase with strength ($M(\epsilon_s) = 0.07$, 95% HDI = $[-0.11, 0.26]$) or their interaction ($M(\epsilon_{s\times w}) = -0.03$, 95% HDI = $[-0.09, 0.02]$).

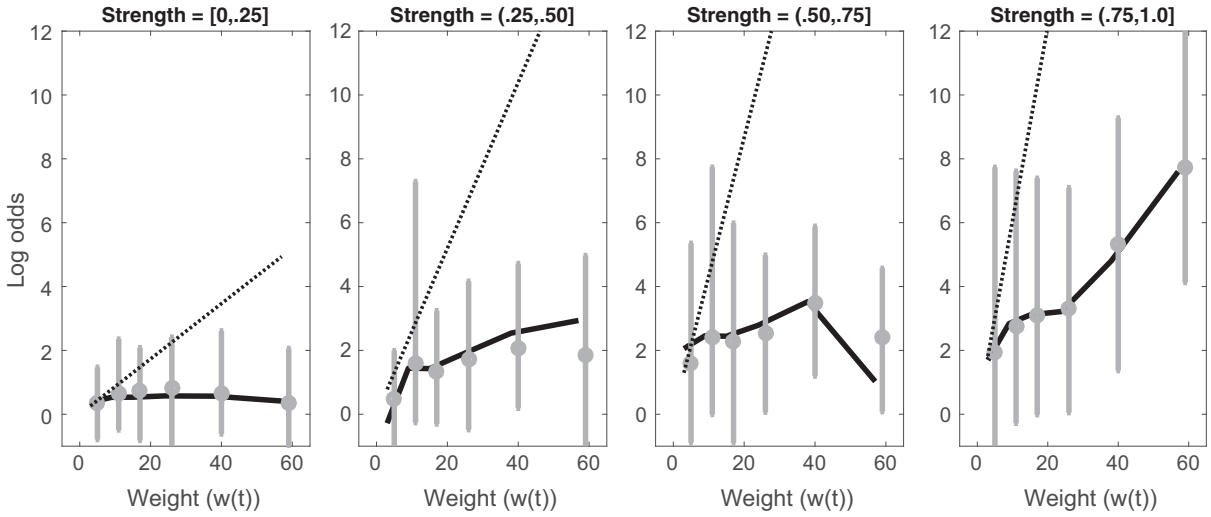

Fig. 3. Mean log odds judgments in the data (solid black lines) and posterior mean and 95% HDI of model predictions (gray) for log odds transformed confidence judgments across strength and weight manipulations. Also shown are the true log odds of the hypotheses for each condition, adjusted for the discriminability $d' = 2$ (dotted black lines). Note that strength and weight were taken directly from the stimuli and therefore vary semi-continuously. Therefore, strength and weight are sorted into categories. Strength groups are given at the top of each plot, and the six weight groups (x values) represent weight values categorized into 0–6, 6–12, 12–18, 18–30, 30–45, and 45+. Note that there were very few high strength - high weight conditions (rightmost points), so these estimates have substantial uncertainty associated with them.

Table 1
Group level mean estimates [95% highest density interval] of standardized linear coefficients predicting diffusion model parameters from strength & weight.

Parameter	Intercept	Strength	Weight	Strength × weight
Drift (μ)	1.28 $[1.20, 1.37]$	0.85 $[0.77, 0.92]$	0.48 $[0.42, 0.55]$	0.37 $[0.31, 0.42]$
Threshold (θ)	3.35 $[3.07, 3.68]$	0.62 $[0.37, 0.85]$	0.11 $[0.00, 0.25]$	0.05 $[-0.05, 0.15]$
Noise (σ^2)	Set to 1	0.11 $[0.02, 0.20]$	0.23 $[0.19, 0.28]$	0.08 $[0.04, 0.12]$

Note. The drift is a function of δ coefficients, threshold is a function of ζ coefficients, and noise is a function of ϵ coefficients.

Table 2
Group level mean estimates [95% highest density interval] of standardized linear coefficients predicting confidence model parameters from strength and weight levels.

Parameter	Intercept	Strength	Weight	Strength × weight
Drift (μ)	0.03 $[-0.17, 0.23]$	0.37 $[0.27, 0.48]$	0.12 $[0.05, 0.20]$	0.11 $[0.07, 0.15]$
Noise (σ^2)	Set to 1	0.24 $[-0.00, 0.47]$	0.49 $[0.09, 0.88]$	0.53 $[0.32, 0.76]$

Note. Drift is a function of δ coefficients and noise is a function of ϵ coefficients.

estimates varied based on the quantity they were asked to estimate – Fig. 4 shows participants' mean estimates of the number of dots (with 95% HDIs) based on the actual number of corresponding dots.

We also fit a hierarchical Bayesian linear model predicting the intercept, slope, and standard deviation of the dot number estimates based on the actual number shown. The (non-standardized) mean group-level estimates are credibly unbiased (mean intercept b_0 = 0.49, 95% HDI = $[-0.06, 1.06]$; mean slope b_1 = 1.01, 95% HDI = $[0.91, 1.12]$) and even trended slightly high, so underestimation of the number of dots is not a plausible explanation for under-emphasis of weight. Instead, it seems that participants are able to store the information presented to them with reasonable accuracy, making it unlikely that failures to use strength and weight information optimally are due to perceptual errors.

The effect of the true number of dots on the *variance* of number estimates, however, suggests that the distribution of these estimates was wider with the number of dots shown (b_σ = 0.50, 95% HDI = $[0.04, 0.99]$). Consistent with the noise parameter estimates in choice and confidence tasks as well as the results from previous numeracy studies, this change in the variability of responses indicates that our participants' precision decreased as they estimated larger numerical magnitudes.

4. Discussion

The most apparent result of our studies is that participants' decisions and judgments were affected more by changes in strength than weight of information. This was reflected in the effect of experimental manipulations of these two factors on the estimated drift rates – across tasks, the best estimates indicate that manipulations of strength have at least 1.5× the effect on drift rate relative to manipulations of weight. These findings violate the proportional representation assumption made by sequential sampling models of the evidence accumulation process.

As such, we replicated the main findings of Griffin and Tversky (1992) in a perceptual task and extended it to cover dynamic judgments as well as decisions. This comes with the additional caveat that the greater effect of strength led to under-confidence in high-weight, low-strength conditions. While we did not find direct evidence for overconfidence in low-weight, high-strength conditions as Griffin and Tversky did, the slope of the estimates and data in Fig. 3 suggests that this would be the case if we were to force lower-weight trials. Overconfidence would also likely arise if we manipulated other sources of class-based information. For example, people's insensitivity to the base rate of green-dominant or pink-dominant dots (base rate neglect; Bar-Hillel, 1980; Griffin & Tversky, 1992; Kahneman & Tversky, 1973) could produce overconfidence based on samples' representativeness. Similarly, manipulating the discriminability of the two hypotheses [d'] by adjusting the proportions of each color to e.g. $\frac{3}{4}$ or $\frac{8}{15}$ could potentially bump participants' judgments and decisions around relative to representations based on the true log odds (Wallsten, 1996).

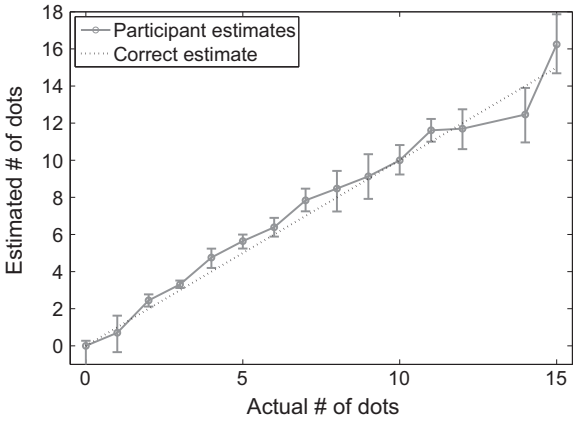

Fig. 4. Mean estimated number of green, pink, or total dots (y) as a function of the actual number displayed (x) in the numerical estimation task. Dots indicate the group level mean of participants' estimates, and the error bars correspond to the 95% HDI of these estimates. The diagonal dotted black line indicates the correct estimates, corresponding to a perfect match between estimated and true number of dots on the screen.

On top of replicating previous findings in a perceptual task, extending them to dynamic decisions, and relating our results to models of the decision-making process, we also investigated the source of these effects. Past explanations of judgments being more affected by strength than weight have been in terms of a dual-process framework (Brenner et al., 2012). According to this framework, the effect of strength is due to the dominance of a heuristic decision system that primarily examines case-based information regarding sample proportion (System 1). The smaller impact of weight would then be attributable to a second information processing system that focuses more on class-based information (System 2). This second system would allow for better adjustment of log odds judgments relative to the true posterior probabilities, but be invoked only later during trials – therefore, it could explain why judgments in low-strength and low-weight trials (which tended to result in longer presentation times because it took participants longer to gather sufficient information) tended to be closer to the true probabilities. However, there are two slight curiosities here: first, we showed via the number estimation task that participants could accurately though not precisely assess the number of dots presented to them. This would suggest that they paid attention to weight information but just didn't use it to make their judgments, a phenomenon which is odd to attribute to a System 1 information processing pattern. Second, the task here was a dynamic perceptual task, as opposed to the higher-level cognitive tasks that dual-process theories typically explain. To account for these data with a dual process explanation, one would have to claim that both systems exist on a fundamental perceptual level. Each of these is possible, but one could envision a simpler single-process explanation that accounts for choice, confidence and response time quantitatively.

Alternatively, the differential effects of the strength versus weight of evidence on perceptual judgment and choice may have a more fundamental source (for analogous results with other phenomena see Pleskac & Busemeyer, 2010; Trueblood, Brown, Heathcote, & Busemeyer, 2013; Zeigenfuse, Pleskac, & Liu, 2014). Our results suggest this may reside in participants' ability to translate the raw strength and weight information they experienced into appropriate evidence representations. The results of our choice, judgment, and numerical estimation tasks all suggest that the precision of participants' estimates of larger magnitudes (in terms of weight) decreased. This result is consistent with findings that people are less able to discriminate via absolute differences between large numbers relative to small numbers (Feigenson et al., 2004; Longo & Lourenco, 2007). This results in a concave sensitivity function for number representations reflecting Weber's law. If evidence representations rely on this sensitivity function, this would explain why higher weight results in progressively greater underconfidence (see Fig. 3). However, the ultimate source of this insensitivity remains an open question – it may be an adaptation to the low prevalence of high-weight information or perhaps to the greater accessibility of strength information.

In any case, it seems that the accumulation of evidence for making judgments and decisions is not a true Bayesian updating process, nor does it proportionally represent strength and weight. However, the weaker assumption that evidence is monotonically related to strength and weight does seem to hold. The drift rate seems to correspond to some combination of strength and weight information, just not the "correct" combination. Our hierarchical linear model predicting drift from these manipulations provided a more than reasonable account of accuracy, response time, and confidence judgments (see Figs. 2 and 3), suggesting that there is at least one function mapping strength and weight onto evidence that can produce our results with high fidelity.

Another curious result is the finding that strength and weight manipulations in the choice task seemed to affect the thresholds that participants set for their decisions. However, strength and weight were not known to participants before the start of a trial. This means that thresholds must be changing somehow while participants are accumulating evidence. Our results suggest that thresholds increase with higher strength and weight, which could be construed in two ways. One construal would be that participants are willing to set higher thresholds when they know they will be met. In this case, participants would simply adopt stricter criteria when they knew they were receiving high quality/quantity information. An alternative construal is that participants in low strength and weight conditions lowered their thresholds. Since the accumulation process would be slower in these conditions, it could be the case that thresholds are collapsing over time such that low strength and low weight conditions hit the decision bounds when they have already partially collapsed. Therefore, this result could likely be explained by a choice process with collapsing bounds, as other authors have suggested (Bowman, Kording, & Gottfried, 2012; Drugowitsch, Moreno-Bote, Churchland, Shadlen, & Pouget, 2012; Ratcliff & Frank, 2012).

4.1. Concluding remarks

In this paper, we examined how people update their beliefs based on the strength (sample proportion) and weight (sample size) of incoming information. In particular, we expanded on previous work by examining the effect of manipulations of these factors on both choices and judgments regarding dynamic perceptual stimuli. Our results suggested that strength contributed much more to evidence representations underlying both decisions and judgments. This indicates a violation of the optimal Bayesian updating mechanism which is thought to underlie evidence accumulation as well as the weaker proportional representation that suggests that these representations noisily mirror the characteristics of the stimulus.

It seems that drift-diffusion and similar evidence accumulation models can still offer a good account of behavioral data, but they must be modified to move away from the Bayesian updating, proportional representation, and static threshold assumptions that have been inherited from early models of this process. We have taken some steps in this paper by providing models which set drift and threshold based on strength and weight manipulations, but further efforts on identifying the source of the strength-weight gap will help us better construct models of how stimulus information maps onto evidence representations used to make judgments and decisions.

Appendix A. Models with absolute stopping rules

When we considered relative stopping rules we defined strength as a linear transformation of the proportion of samples which favored option A at time t:

$$s(t) = \frac{A(t) - B(t)}{A(t) + B(t)} = 2 \cdot \frac{A(t)}{A(t) + B(t)} - 1 \tag{7}$$

However, for absolute stopping rules it is more useful to use the original sample proportion $sp(t)$.

$$sp(t) = \frac{A(t)}{A(t) + B(t)} = \frac{s(t) + 1}{2} \tag{8}$$

If a sequential sampling model were to use an absolute stopping rule, where an answer is given as soon as either A or B gains enough samples, then there are 2 accumulators to consider. Each of their positions is given as

$$\begin{aligned} A(t) &= (A(t) + B(t)) \cdot \frac{(2 \cdot \frac{A(t)}{A(t)+B(t)} - 1) + 1}{2} \\ &= w(t) \cdot sp(t) \\ B(t) &= (A(t) + B(t)) \cdot \left(1 - \frac{(2 \cdot \frac{A(t)}{A(t)+B(t)} - 1) + 1}{2}\right) \\ &= w(t) \cdot (1 - sp(t)) \end{aligned} \tag{9}$$

In essence, the accumulator for A is the weight times the original sample proportion instantiation of strength (with its linear transformation undone), and the accumulator for B is the weight times (1− sample proportion). The drift rates for these accumulators can be found by taking the derivative, which will result in the rate of change of weight times the strength as in Eq. (6). The important thing to note is that each accumulator is still an even function of strength and weight — neither one emphasizes one over the other. Therefore, essentially all of the predictions we discuss regarding relative stopping rule models will also hold for absolute stopping rule models.

Appendix B. Model details, data, and code

The model code, raw data, and MATLAB scripts for analyses are available on the Open Science Framework at https://osf.io/ba5c7/.

For all models presented, vague priors consistent with those used by Wabersich and Vandekerckhove (2014) were used for each parameter so as to let the data have maximal influence on the posterior estimates. Choice and confidence model parameters were estimated using 8 parallel chains. Each chain was comprised of 1000 burn-in steps (unrecorded samples to allow the chain to reach the reasonable parameter space) and 5000 samples. Preliminary analyses confirmed that all chains converged.

B.1. Choice model

The JAGS code for the diffusion model we used is given below. The inputs to the model are nData (the number of data points), nSubjects (the number of participants), strength (standardized value of strength for each trial), weight (standardized value of weight for each trial), subject (number indicating which participant the data point corresponds to), and y (response times, with incorrect responses coded as negative response times). It requires the dweiner package from Wabersich and Vandekerckhove (2014) in order to run.

```
model {
  for( i in 1:nData ) {
    y[i] ~ idwiener(thresh[i], tau[subject[i]], .5,
  drift[i])
    drift[i] <- (d0[subject[i]] + d1[subject[i]]*
  strength[i] + d2[subject[i]]*weight[i] + d12
  [subject[i]]*strength[i]*weight[i])/noise[i]
    thresh[i] <- (a0[subject[i]] + a1[subject[i]]*
  strength[i] + a2[subject[i]]*weight[i] + a12
  [subject[i]]*strength[i]*weight[i])/noise[i]
    noise[i] <- exp(lognoise[i])
    lognoise[i] <- n1[subject[i]]*strength[i] + n2
  [subject[i]]*weight[i] + n12[subject[i]]*
  strength[i]*weight[i]
}
# Priors

Mt ~ dnorm(0, .0001)
Pt ~ dgamma(.001, .001)

for( s in 1:nSubjects ) {
      tau[s] ~ dnorm(Mt, Pt)

      d0[s] ~ dnorm(Md0,Pd0)
      d1[s] ~ dnorm(Md1,Pd1)
      d2[s] ~ dnorm(Md2,Pd2)
      d12[s] ~ dnorm(Md12,Pd12)

      a0[s] ~ dnorm(Ma0,Pa0)
      a1[s] ~ dnorm(Ma1,Pa1)
      a2[s] ~ dnorm(Ma2,Pa2)
      a12[s] ~ dnorm(Ma12,Pa12)

      n1[s] ~ dnorm(Mn1,Pn1)
      n2[s] ~ dnorm(Mn2,Pn2)
      n12[s] ~ dnorm(Mn12,Pn12)
  }

  Md0 ~ dnorm(0, .0001)
  Md1 ~ dnorm(0, .0001)
  Md2 ~ dnorm(0, .0001)
  Md12 ~ dnorm(0, .0001)

  Ma0 ~ dnorm(0, .0001)
  Ma1 ~ dnorm(0, .0001)
  Ma2 ~ dnorm(0, .0001)
  Ma12 ~ dnorm(0, .0001)

  Mn1 ~ dnorm(0, .0001)
  Mn2 ~ dnorm(0, .0001)
  Mn12 ~ dnorm(0, .0001)

  Pd0 ~ dgamma(.001, .001)
  Pd1 ~ dgamma(.001, .001)
  Pd2 ~ dgamma(.001, .001)
  Pd12 ~ dgamma(.001, .001)

  Pa0 ~ dgamma(.001, .001)
  Pa1 ~ dgamma(.001, .001)
  Pa2 ~ dgamma(.001, .001)
  Pa12 ~ dgamma(.001, .001)

  Pn1 ~ dgamma(.001, .001)
  Pn2 ~ dgamma(.001, .001)
  Pn12 ~ dgamma(.001, .001)
 }
```

B.2. Confidence model code

Confidence judgments of 0% and 100% were transformed to 0.1% and 99.9% before being turned into log odds (to avoid divisions by zero). The JAGS code for this model is presented below. Note that the confidence model uses a t distribution to describe the log odds — this is done to provide the heavy-tailed shape of the data due arising from the large number of estimates near 0% and 100%.

As in the choice model, inputs to the model are nData (the number of data points), nSubjects (the number of participants), strength (standardized value of observed strength for each trial), weight (standardized value of observed weight for each trial), subject (number indicating which participant the data point corresponds to), and y (log odds judgments).

```
model {
for( i in 1: nData) {
    y[i] ~ dt(drift[i],tau[i],4)

    drift[i] <- d0[subject[i]] + dl[subject[i]]*
  strength[i] + d2[subject[i]]*weight[i] + dl2
  [subject[i]]*strength[i]*weight[i]
    tau[i] <- 1/pow(exp(logsig[i]), 2 )
    logsig[i] <- s0 + sl[subject[i]]*strength[i] +
  s2[subject[i]]*weight[i] + sl2[subject[i]]*
  strength[i]*weight[i]

 }
# Priors

for( n in 1:nSubjects ) {
    d0[n] ~ dnorm(Md0, Pd0)
    dl[n] ~ dnorm(Mdl, Pdl)
    d2[n] ~ dnorm(Md2, Pd2)
    dl2[n] ~ dnorm(Mdl2, Pdl2)

    sl[n] ~ dnorm(Msl,Psl)
    s2[n] ~ dnorm(Ms2,Ps2)
    sl2[n] ~ dnorm(Msl2,Psl2)

}

s0 ~ dgamma(.001, .001)

# Priors

  Md0 ~ dnorm(0, .0001)
  Mdl ~ dnorm(0, .0001)
  Md2 ~ dnorm(0, .0001)
  Mdl2 ~ dnorm(0, .0001)

  Msl ~ dnorm(0, .0001)
  Ms2 ~ dnorm(0, .0001)
  Msl2 ~ dnorm(0, .0001)

  Pd0 ~ dgamma(.001, .001)
  Pdl ~ dgamma(.001, .001)
  Pd2 ~ dgamma(.001, .001)
  Pdl2 ~ dgamma(.001, .001)

  Psl ~ dgamma(.001, .001)
  Ps2 ~ dgamma(.001, .001)
  Psl2 ~ dgamma(.001, .001)

}
```

References

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

Beck, J. M., Ma, W. J., Kiani, R., Hanks, T., Churchland, A. K., Roitman, J., & Pouget, A. (2008). Probabilistic population codes for Bayesian decision making. *Neuron, 60*(6), 1142–1152. http://dx.doi.org/10.1016/j.neuron.2008.09.021.

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. http://dx.doi.org/10.1037/0033-295X.113.4.700.

Bowman, N. E., Kording, K. P., & Gottfried, J. A. (2012). Temporal integration of olfactory perceptual evidence in human orbitofrontal cortex. *Neuron, 75*(5), 916–927. http://dx.doi.org/10.1016/j.neuron.2012.06.035.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*, 433–436. http://dx.doi.org/10.1163/156856897X00357.

Brenner, L. A., Griffin, D. W., & Koehler, D. J. (2012). A case-based model of probability and pricing judgments: Biases in buying and selling uncertainty. *Management Science, 58*(1), 159–178. http://dx.doi.org/10.1287/mnsc.1110.1429.

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review, 100*(3), 432–459. http://dx.doi.org/10.1037//0033-295X.100.3.432.

Donkin, C., Brown, S. D., & Heathcote, A. (2009). The overconstraint of response time models: Rethinking the scaling problem. *Psychonomic Bulletin & Review, 16*(6), 1129–1135. http://dx.doi.org/10.3758/PBR.16.6.1129.

Drugowitsch, J., Moreno-Bote, R., Churchland, A. K., Shadlen, M. N., & Pouget, A. (2012). The cost of accumulating evidence in perceptual decision making. *The Journal of Neuroscience, 32*(11), 3612–3628. http://dx.doi.org/10.1523/JNEUROSCI.4010-11.2012.

Edwards, W. (1965). Optimal strategies for seeking information: Models for statistics, choice reaction times, and human information processing. *Journal of Mathematical Psychology, 2*(2), 312–329. http://dx.doi.org/10.1016/0022-2496(65)90007-6.

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences, 8*(7), 307–314. http://dx.doi.org/10.1016/j.tics.2004.05.002.

Gallistel, C. R., & Gelman, R. (2000). Non-verbal numerical cognition: From reals to integers. *Trends in Cognitive Sciences, 4*(2), 59–65. http://dx.doi.org/10.1016/S1364-6613(99)01424-2.

Gold, J. I., & Shadlen, M. N. (2001). Neural computations that underlie decisions about sensory stimuli. *Trends in Cognitive Sciences, 5*(1), 10–16. http://dx.doi.org/10.1016/S1364-6613(00)01567-9.

Green, D., & Swets, J. (1966). *Signal detection and psychophysics*.New York, NY: Wiley & Sons, Inc..

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology, 24*(3), 411–435. http://dx.doi.org/10.1016/0010-0285(92)90013-R.

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science, 15*(8), 534–539. http://dx.doi.org/10.1111/j.0956-7976.2004.00715.x.

Hertwig, R., & Erev, I. (2009). The description-experience gap in risky choice. *Trends in Cognitive Sciences, 13*, 517–523. http://dx.doi.org/10.1016/j.tics.2009.09.004.

Kahneman, D. (2003). A perspective on judgment and choice – Mapping bounded rationality. *American Psychologist, 58*(9), 697–720. http://dx.doi.org/10.1037/0003-066x.58.9.697.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*(4), 237–251. http://dx.doi.org/10.1037/h0034747.

Kepecs, A., Uchida, N., Zariwala, H. A., & Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature, 455*(7210), 227–231. http://dx.doi.org/10.1038/nature07200.

Kiani, R., & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science, 324*(5928), 759–764. http://dx.doi.org/10.1126/science.1169405.

Kira, S., Yang, T., & Shadlen, M. N. (2015). A neural implementation of Wald's sequential probability ratio test. *Neuron, 85*(4), 861–873. http://dx.doi.org/10.1016/j.neuron.2015.01.007.

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in Psychtoolbox-3. *Perception, 36*(14), 1.

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences, 27*(12), 712–719. http://dx.doi.org/10.1016/j.tins.2004.10.007.

Krajbich, I., Lu, D., Camerer, C., & Rangel, A. (2012). The attentional drift-diffusion model extends to simple purchasing decisions. *Frontiers in Psychology, 3*. http://dx.doi.org/10.3389/fpsyg.2012.00193.

Kruschke, J. (2010). *Doing Bayesian data analysis: A tutorial introduction with R*. Academic Press.

Laming, D. R. J. (1968). *Information theory of choice-reaction times*. Academic Press.

Lee, M. D., & Wagenmakers, E. J. (2013). *Bayesian modeling for cognitive science: A practical course*.New York, NY: Cambridge University Press.

Link, S., & Heath, R. (1975). A sequential theory of psychological discrimination. *Psychometrika, 40*(1), 77–105. http://dx.doi.org/10.1007/BF02291481.

Longo, M. R., & Lourenco, S. F. (2007). Spatial attention and the mental number line: Evidence for characteristic biases and compression. *Neuropsychologia, 45*(7), 1400–1407. http://dx.doi.org/10.1016/j.neuropsychologia.2006.11.002.

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience, 9*(11), 1432–1438. http://dx.doi.org/10.1038/nn1790.

Merkle, E. C., & Van Zandt, T. (2006). An application of the Poisson race model to confidence calibration. *Journal of Experimental Psychology: General, 135*(3), 391–408. http://dx.doi.org/10.1037/0096-3445.135.3.391.

Meyniel, F., Sigman, M., & Mainen, Z. F. (2015). Confidence as Bayesian probability: From neural origins to behavior. *Neuron, 88*(1), 78–92. http://dx.doi.org/10.1016/j.neuron.2015.09.039.

Moran, R., Teodorescu, A. R., & Usher, M. (2015). Post choice information integration as a causal determinant of confidence: Novel data and a computational account. *Cognitive Psychology, 78*, 99–147. http://dx.doi.org/10.1016/j.cogpsych.2015.01.002.

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review, 104*(2), 266–300. doi:http://dx.doi.org/10.1037/0033-295X.104.2.266.

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*(5), 376–404. http://dx.doi.org/10.1167/5.5.1.

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review, 117*(3), 864–901. http://dx.doi.org/10.1037/A0019737.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*(2), 59–108. http://dx.doi.org/10.1037/0033-295X.85.2.59.

Ratcliff, R. (2014). Measuring psychometric functions with the diffusion model. *Journal of Experimental Psychology: Human Perception and Performance, 40*(2), 870–888. http://dx.doi.org/10.1037/a0034954.

Ratcliff, R., & Frank, M. J. (2012). Reinforcement-based decision making in corticostriatal circuits: Mutual constraints by neurocomputational and diffusion models. *Neural Computation, 24*(5), 1186–1229. http://dx.doi.org/10.1162/NECO-a-00270.

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

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

Ratcliff, R., & Starns, J. J. (2009). Modeling confidence and response time in recognition memory. *Psychological Review, 116*(1), 59–83. http://dx.doi.org/10.1037/A0014086.

Ratcliff, R., & Starns, J. J. (2013). Modeling confidence judgments, response times, and multiple choices in decision making: Recognition memory and motion discrimination. *Psychological Review, 120*(3), 697–719. http://dx.doi.org/10.1037/a0033152.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*(1), 3–22. http://dx.doi.org/10.1037//0033-2909.119.1.3.

Smith, P. L., & Van Zandt, T. (2000). Time-dependent Poisson counter models of response latency in simple judgment. *British Journal of Mathematical and Statistical Psychology, 53*(2), 293–315. http://dx.doi.org/10.1348/000711000159349.

Stone, M. (1960). Models for choice-reaction time. *Psychometrika, 25*(3), 251–260. http://dx.doi.org/10.1007/BF02289729.

Trueblood, J. S., Brown, S. D., Heathcote, A., & Busemeyer, J. R. (2013). Not just for consumers context effects are fundamental to decision making. *Psychological Science, 24*(6), 901–908. http://dx.doi.org/10.1177/0956797612464241.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review, 108*(3), 550–592. http://dx.doi.org/10.1037/0033-295X.108.3.550.

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical diffusion models for two-choice response times. *Psychological Methods, 16*(1), 44–62. http://dx.doi.org/10.1037/a0021765.

Van den Berg, R., Anandalingam, K., Zylberberg, A., Kiani, R., Shadlen, M. N., & Wolpert, D. M. (2016). A common mechanism underlies changes of mind about decisions and confidence. *Elife, 5*, e12192. http://dx.doi.org/10.7554/eLife.12192.

Vickers, D. (1979). *Decision processes in visual perception*.London: Academic Press.

Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. *Memory & Cognition, 32*(7), 1206–1220. http://dx.doi.org/10.3758/BF03196893.

Wabersich, D., & Vandekerckhove, J. (2014). Extending JAGS: A tutorial on adding custom distributions to JAGS (with a diffusion model example). *Behavior Research Methods, 46*(1), 15–28. http://dx.doi.org/10.3758/s13428-013-0369-3.

Wald, A., & Wolfowitz, J. (1949). Bayes solutions of sequential decision problems. *Proceedings of the National Academy of Sciences of the United States of America, 35*(2), 82–99. http://dx.doi.org/10.1214/aoms/1177729887.

Wallsten, T. S. (1996). An analysis of judgment research analyses. *Organizational Behavior & Human Decision Processes, 65*(3), 220–226. http://dx.doi.org/10.1006/obhd.1996.0022.

Yu, S., Pleskac, T. J., & Zeigenfuse, M. D. (2015). Dynamics of postdecisional processing of confidence. *Journal of Experimental Psychology: General, 144*(2), 489–510. http://dx.doi.org/10.1037/xge0000062.

Zeigenfuse, M. D., Pleskac, T. J., & Liu, T. S. (2014). Rapid decisions from experience. *Cognition, 131*(2), 181–194. http://dx.doi.org/10.1016/j.cognition.2013.12.012.