Internal and External Sources of Variability in Perceptual Decision-Making

Roger Ratcliff\textsuperscript{1*}, Chelsea Voskuilen\textsuperscript{1}, and Gail McKoon\textsuperscript{1}

The Ohio State University

\textsuperscript{1}The Ohio State University,
Department of Psychology,
Columbus, OH, 43210 USA

(614) 292-7916

*Address correspondence to:
Roger Ratcliff
email ratcliff.22@osu.edu
Abstract

It is important to identify sources of variability in processing to understand decision-making in perception and cognition. There is a distinction between internal and external variability in processing and double-pass experiments have been used to estimate their relative contributions. In these and our experiments, exact perceptual stimuli are repeated later in testing and agreement on the two trials is examined to see if it is greater than chance. In recent research in modeling decision processes, some models implement only (internal) variability in the decision process while others explicitly represent multiple sources of variability. We describe five perceptual double-pass experiments which show greater than chance agreement which is inconsistent with models that assume internal variability alone. Estimates of total trial-to-trial variability in the evidence accumulation (drift) rate (the decision-relevant stimulus information) were estimated from fits of the standard diffusion decision-making model to the data. The double-pass procedure provided estimates of how much of this total variability was systematic and dependent on the stimulus. These results provide the first behavioral evidence independent of model fits for trial-to-trial variability in drift rate in tasks used in examining perceptual decision making.

Keywords: double-pass procedure; diffusion decision model; response time and accuracy; trial to trial variability
Introduction

The notion that human information processing is noisy is fundamental to psychology and neuroscience. Green (1964) and Swets, Shipley, McKey, and Green (1959) made a distinction between two kinds of noise. Internal noise was said to come from moment-to-moment variability in the state of the processing system and external noise from variability in the representations encoded from stimuli. In experimental studies, external noise has been added to stimuli by, for example, adding acoustic noise to acoustic stimuli (the Green and Swets et al. studies) or random pixel noise to visual stimuli (e.g., Gabor patches, Lu & Dosher, 2008). Another source of external noise arises from variability in the individual instances that are tested from a specified class of stimuli.

Most previous investigations of external and internal noise have used only accuracy as the dependent variable and their results have been interpreted in terms of signal detection theory, in which all sources of noise are combined. Instead, sequential sampling models (Ratcliff & Smith, 2004) are intended to constrain interpretations of data with both response times (RTs) and accuracy and to separate sources of noise. Some sequential sampling models implement both external and internal sources of noise (e.g., Ratcliff; 1978; Ratcliff & McKoon, 2008; Vandekerckhove & Tuerlinckx, 2008; Voss & Voss, 2008; Wiecki et al., 2013) but others only internal noise (Churchland, Kiani, & Shadlen, 2008; Deneve, 2012; Ditterich, 2006a, 2008; Drugowitsch, Moreno-Bote, Churchland, Shadlen, & Pouget, 2012; Hanks, Mazurek, Kiani, Hopp, & Shadlen, 2011; Kiani, Corthell, & Shadlen, 2014; Palmer, Huk, & Shadlen, 2005; Usher & McClelland, 2001; Zhang, Lee, Vandekerckhove, Maris, & Wagenmakers, 2014).

The question, then, is whether current models can explain accuracy and RT data with only internal noise or whether the external noise, or variation between stimulus exemplars, is also required—even when literal noise is not added to the stimulus. In this article, we provide direct evidence that external noise is, in fact, required to explain the data from five simple two-choice decision tasks with perceptual and cognitive stimuli. We use the sequential sampling diffusion model developed by Ratcliff (Ratcliff, 1978; Ratcliff & McKoon, 2008) to separate
sources of variability.

In the two-choice diffusion model (Figure 1A), information from the encoded representation of a stimulus accumulates from a starting point toward one of two boundaries and when a boundary is reached, a response is executed. The rate of accumulation (drift rate) is determined by the quality of the representation with respect to the information needed to make a decision. The accumulation process is noisy (the jagged lines in the figure) so that for a given drift rate, the process will reach boundaries at different times, giving distributions of RTs, and sometimes reach the wrong boundary, giving errors. This is an internal source of variability that we label "within-trial" variability. There are also sources of internal variability that play out across trials. For a given stimulus, the representation encoded from it, that is, its drift rate, can fluctuate across trials as a function of, for example, attention, motivation, or varying levels of fatigue. For the same reasons, the location of the starting point can fluctuate across trials and so can the time taken up by processes outside the decision process itself, which include stimulus encoding time, the time to extract decision-related information from the encoded representation, and response execution time.

Noise from these internal sources comes from the processing system itself, moment-to-moment variability within a trial between the starting point of accumulation and the boundaries and moment-to-moment variability across trials in drift rate, starting point, and nondecision time. External variability is instead linked to stimuli. We use the numerosity discrimination experiment described below to illustrate it (Figure 2A): Subjects were shown arrays of asterisks and for each, they were asked to decide whether the number of asterisks in it was more or less than 50. The difficulty of the decision was manipulated by the number of asterisks, for example, a stimulus of 45 asterisks is more difficult than a stimulus of 35 asterisks. For each number of asterisks, the configuration of the asterisks changes from trial to trial. Sometimes 35 asterisks might be grouped in the center of the array, or they might be widely scattered, or they might be mostly in the upper right corner, and so on, and some configurations might be more obviously different from 50 than others. The question is whether this variability from one instance of 35 asterisks to another affects the representation that is encoded from the stimuli (and therefore
affects drift rate). If so, it is variability that comes from the stimulus, that is it is external variability; it does not come from internal states of the processing system. Many of the implemented models cited above assume that this external variability is not necessary to explain data.

We measured the contributions of internal and external noise to performance with a double pass manipulation (Green, 1964; see also Burgess & Colborne, 1988; Cabrera et al., 2015; Gold et al., 1999; Lu & Dosher, 2008; 2014). Over the trials of an experiment, exactly the same stimulus (i.e., the same number of asterisks displayed in the same configuration) is presented twice with some large number of trials intervening.

Without the external variability that we just defined, the encoded representation of 35 asterisks would be subject only to trial-to-trial internal noise. Drift rates for the two presentations would be identical to the mean for 35 asterisks and responses to them would be independent. With external variability, drift rates for the two presentations can be systematically different from the mean, larger for easier configurations than difficult ones. Differences in RTs and choices would come from the combination of external and across-trial internal noise. If the responses are correlated, then the correlation must come from external noise. If responses are easier for 35 asterisks when they are widely dispersed than when they are clustered, they would be easier for both presentations. These relationships are illustrated in Figure 1B and 1C. $v$ is the mean drift rate for a given number of asterisks (e.g., 35). The blue dotted lines represent a configuration that has a lower drift rate and the red dashed lines one that has higher drift rate. Without across-trial internal variability, the drift rates on the first and second presentations would be the same. Internal variability gives differences, shown by the two red dashed lines and the two blue dotted lines in Figure 1A. Figure 1B shows their drift rates input to the diffusion decision process.

Figure 1D shows how double pass accuracy data can be displayed (e.g., Burgess & Colborne, 1988; Lu & Dosher, 2008): accuracy is plotted against the probability that the response choice on the two presentations is the same (called their probability of agreement). For
each of the curves in the figure, accuracy begins at chance and rises to ceiling. (Different levels of accuracy would typically come from conditions in an experiment that differed in difficulty.) Each curve is labeled with a correlation between the two responses. For the furthest left function, the correlation is zero (i.e., the responses are independent) and the probability of agreement is $p*p+(1-p)*(1-p)$ (e.g., for $p=0.7$, agreement=0.58). The further right functions show increasing correlations and their increasing probabilities of agreement. Examples of how accuracy and agreement are calculated are shown in Appendix B.

Experiments 1-5

The tasks were all two-choice tasks and they are illustrated in Figure 2. The procedures were the same as in previously published experiments. Full details are given in Appendix A. Either the first 90 or 96 trials (depending on the experiment) were exactly repeated, in the same order, in a second block. (We replicated one of the experiments, Experiment 1, with the stimuli in random order in the second block relative to the first block and the results were almost identical.)

In the numerosity discrimination task (Ratcliff, 2014; Ratcliff, Thapar, & McKoon, 2001, 2010; Ratcliff, Thompson, & McKoon, 2015), Experiment 1, subjects decided whether the number of asterisks in a 10x10 array was greater or less than 50. Difficulty was manipulated with numbers closer and further from 50.

In the letter discrimination task (Ratcliff & Smith, 2010; Thapar, Ratcliff, & McKoon, 2003), Experiment 2, subjects decided whether a centrally displayed letter was one of two choices that were displayed to the left and right of a fixation point. Difficulty was manipulated by the time a letter was displayed, 10, 20, or 30 ms, before it was masked by randomly oriented lines.

In the motion discrimination task (Britten et al., 1992; Palmer et al., 2005; Ratcliff & McKoon, 2008; Roitman & Shadlen, 2002; Shadlen & Newsome, 2001; Salzman et al., 1992), Experiment 3, a stimulus was composed of dots in a circular window displayed for 400 ms. On
each trial, some proportion of the dots moved in one direction, either to the left or right, and the rest moved in random directions. Subjects were asked to decide whether the direction of the coherently moving dots was to the left or right. Stimulus difficulty was varied via the proportion of dots moving in the same direction, 0.10, 0.15, or 0.20.

In the static brightness discrimination task (Ratcliff, 2002; Ratcliff & Smith, 2010; Ratcliff, Thapar & McKoon, 2003), Experiment 4, a 64x64 array of black and white pixels was displayed for 100 ms and then masked by checkerboard pixel arrays. Subjects decided whether there were more white pixels or more black ones. Difficulty was manipulated with the proportion of white pixels, 0.43, 0.46, 0.54, or 0.57.

The dynamic brightness task (Ratcliff & Smith, 2010), Experiment 5, was the same as the static one except that a different array of 60x60 black and white pixels was displayed every 16.67 ms until a response was made. Difficulty was manipulated with the proportion of white pixels, 0.46, 0.48, 0.52, or 0.54.

Results

For all five tasks, responses for the two choices were symmetric (e.g., accuracy and RTs were about the same for left-moving dots as right-moving dots), so we combined correct responses for the two choices and we combined error responses for the two. In the diffusion model analyses, this allowed us to fix the starting point at half the boundary separation. We fit the model to the data for each subject individually with a standard quantile-based method that is described in Appendix C (Ratcliff & Childers, 2015; Ratcliff & Tuerlinckx, 2002). The fit was good, as demonstrated by the matches between predictions and data in the quantile-probability plots (averaged over subjects) shown in Figure 3 and the G-square goodness of fit values shown in Table 1. The plots show the .1, .3, .5, .7, and .9 quantile RTs (the vertical columns of points) plotted against accuracy values with errors on the left and correct responses on the right. The plots show how RT distributions change with accuracy (Ratcliff & McKoon, 2008). The best fitting values of the model's parameters are shown in Table 1.

To examine agreement between repeated tests of stimuli, we simulated choices and RTs
for seven levels of drift rate and seven levels of across-trial variability in drift rate. We used the values that best fit the data for boundary separation, nondecision time, across-trial variability in starting point (equivalent to across-trial variability in boundaries), and across-trial variability in nondecision time. For each of the seven drift rates ($v$) and the seven values of across-trial variability in drift rate ($\eta$), the drift rate for a stimulus was chosen randomly from a normal distribution with mean $v$ and SD $\eta$. Random values of starting point and nondecision time were generated from their distributions. For the second simulated presentation of the stimulus, the same drift rate was used but different random samples were selected for starting point and nondecision time (in other words, the drift rate for the two presentations of the stimulus was the same).

For each combination of drift rate and across-trial variability in drift rate, we generated 20,000 simulated choices and RTs (using the random walk method, Tuerlinckx et al., 2001) and then used these simulated data to give the points for the functions shown in Figures 4A-E. The seven drift rates give the seven levels of accuracy on the y-axis which were joined by the lines and the seven values of across-trial variability in drift rate (shown at the bottom of each function) give the seven levels of agreement across the x-axis (see the calculations of agreement in Appendix B). As across-trial variability in drift rate increases, agreement probability increases. The heavy black line is the one that most nearly corresponds to the value of across-trial variability in drift rate from the best fits of the diffusion model to the experimental accuracy and RT data (Table 1).

For each experiment, there were either two or three conditions that differed in difficulty. We calculated accuracy-agreement values from the data for each condition for each subject then averaged them over subjects, and these are the points marked by squares in the figures. The squares all fall on or very close to one of the accuracy-agreement functions. For the numerosity and letter tasks, the squares fall near the function for which across-trial variability is 0.1, for the dot motion task they fall near the 0.15 function, and for the static and dynamic brightness tasks they fall near 0.05. The differences among the configurations of the stimuli across trials were larger for the numerosity, letter, and dot motion tasks than the dynamic and static brightness
tasks, which explains why their probabilities of agreement were further from chance.

The accuracy-agreement functions and the across-trial variability in drift rate values obtained from fitting the model to the data allow across-trial variability in drift rate to be split into external and internal sources. Internal across-trial variability in drift rate produces differences in drift rate that vary from trial to trial and are not systematically related to the exact stimulus presented (e.g., it depends on moment to moment fluctuations in attention, vigilance, sequential effects, and so on). The differences between the squares and the heavy black lines represent across-trial internal variability. For the motion task, there is little internal variability (the squares lie close to the heavy line). For the other tasks, internal variability is larger (the squares lie farther from the heavy line).

For the simulations above, the drift rate for the first and second presentations of a stimulus was the same; there was no across-trial internal variability. To check that this did not affect our conclusions, we added additional (internal) variability in drift rate across trials for the numerosity discrimination task (Figure 4A). For the functions with $\eta=0$, 0.05, and 0.1, we added normally distributed random variability to the simulated drift rates to make the total $\eta=0.15$. This reduced accuracy but the important result was that the shapes and locations of the curves remained the same (the points on each accuracy-agreement function moved down the curve but not off the curve). In other words, the addition of internal across-trial variability did not change the conclusions.

In addition to accuracy-agreement analyses, we also examined the agreement between RTs for the first and second presentations and found that the model fit them well. We calculated the correlation between the two presentations for each condition, task, and subject and averaged them. The average was small, 0.079, and fell between 0.03 and 0.12 for the tasks. To fit the model to the correlations, we used the values of drift rate and across-trial variability in drift rate from the data (the squares in Figures 4A-E); the resulting correlation was 0.03. There is also trial-to-trial variability in nondecision time. If we assume that half that variability is common across presentations of the same stimulus (e.g., encoding time varies systematically across
stimuli), then the mean correlation predicted by the model would increase to 0.06, which is within the range of those from the data. Figure 4F shows a heat map of 10,000 RTs with a correlation of 0.106 to illustrate what a correlation at the upper end would look like. Thus the agreement between RTs on repeated trials is consistent with the diffusion model predictions.

Discussion

The diffusion model fit the data from the five experiments well, explaining both accuracy and RTs and requiring across-trial variability in drift rates. Simulations based on the best fitting values of the boundary settings, nondecision time, across-trial variability in the starting point, and across-trial variability in nondecision time were used to generate accuracy-agreement functions. These assumed trial-to-trial variability in drift rate, but an identical drift rate between the two simulated presentations of the same stimulus. The empirical accuracy-agreement values lay on functions for which the external variability in drift rates was larger than zero for all five experiments. The diffusion model also accounted for the relatively small correlations between RTs on the first and second presentations.

The four perceptual tasks were chosen because they are typical of perceptual discrimination tasks that have been used in previous studies. For these tasks and the numerosity discrimination task, it is extremely unlikely that subjects could use memory for the first presentation of a stimulus to bias processing on a second presentation 100 trials later for the hundreds of such pairs in the experiments. In contrast, memory for stimuli could be an issue in cognitive tasks such as recognition memory, lexical decision, and memory for pictures. If a bias toward one or the other of the responses on the second presentation was the same as on the first, it could be due to memory for the first and this would be indistinguishable from a bias that was the result of consistent differences in encoding processes, that is, external variability.

Much of the research that has used the double-pass method has taken place in the context of what are called observer models of perceptual processing and in many studies, external noise is added to stimuli (e.g., our Experiment 2) to provide a measure of internal variability as a
function of the level of external noise (Burgess & Colborne, 1988; Cabrera, Lu, & Dosher, 2015; Gold, Bennett, & Sekular, 1999; Green, 1964; Lu & Dosher, 2008; 2014; Swets et al., 1959). When the amount of external noise is small, changes in it have little effect on performance but when the amount is larger, performance drops as the amount of external variability approaches the amount of internal variability. Analyses of the data with signal detection theory for the sources of noise combined produce accuracy and d’ measures. In these models, the signal is transformed in various ways and separate sources of noise are all explicitly represented. For example, in Lu and Dosher (2008), several models were examined including their perceptual template model (Lu & Dosher, 1999) in which transformations of signals include rectification and nonlinear gain control and both additive and multiplicative noise produce variability in processing.

These models have been tested only against accuracy data. If they were integrated with the diffusion model, they could be tested jointly against accuracy and RT data. The encoded representations would give drift rates and the diffusion model would translate them into RTs and accuracy. This would provide a stronger test of the observer models than accuracy alone. However, because the diffusion model already has sources of noise, the sources of noise in the observer models might have to be modified in order to produce an integrated model that was consistent with the diffusion model so as to produce values of accuracy and RTs that fit the data. Previously, there have been only a few models that integrate a model of how perceptual or cognitive stimuli are represented with a diffusion decision model. Some examples are those for perceptual letter matching by Ratcliff (1981), for perceptual stimuli by Smith and Ratcliff (2009), for numerosity stimuli by Ratcliff and McKoon (submitted), and for reinforcement learning by Pederson, Frank, and Biele (2016; although they did not include trial-to-trial variability in model parameters). EEG measures have provided support for trial-to-trial variability in the evidence that drives decision processes (drift rate). Ratcliff, Philiastides, and Sajda (2009) used EEG and behavioral data from a two-choice face-car discrimination task. In their study, a single regressor value was computed from a weighted sum of the EEG data for each of several time windows for each trial. The regressor represents how face-like or car-like
the stimulus was and because it was based only on whether the stimulus was a face or car, it was independent of the behavioral data. The regressor was used to sort the behavioral data for each condition (level of difficulty) into more face-like and more car-like stimuli and the diffusion model was fit to these two halves of the behavioral data. Drift rates were different for the two halves of the data, showing that the EEG measure tapped into trial-to-trial differences in the evidence driving the decision process. Ratcliff, Sederberg, Smith, & Childers (2016) obtained a similar result for recognition memory, a task for which subjects decide whether a test word had or had not appeared on a previously presented list. A regressor derived from EEG signals based only on whether a test item had or had not appeared in the list produced differences in drift rates when the diffusion model was fit to the behavioral data for the two halves (see Amitay et al., 2013, for a related study in auditory perception). However, unlike results from the double-pass procedure, the trial-to-trial variability measured by EEG signals might be across-trial internal variability, external variability, or a combination of the two.

Trial-to-trial variability in drift rate was originally motivated in the diffusion model by the plausible notion that individuals cannot extract identical representations of stimuli across trials (Ratcliff, 1978). This has the byproduct that it explains why errors are often slower than correct responses, something that models without external variability cannot do unless additional assumptions are made. One popular such assumption is that boundaries collapse over the accumulation of evidence and another is that drift rate increases over it (via an urgency signal; e.g., Churchland et al., 2008; Deneve, 2012; Ditterich, 2006a, 2006b; Drugowitsch et al., 2012; Hanks et al., 2011; Kiani et al., 2014). Milosavljevic et al. (2010), Hawkins et al. (2015), and Voskuilen et al. (2016) compared the standard diffusion model (constant boundaries and across-trial variability in drift rates) to models with collapsing boundaries and without across-trial variability in drift rates and found that the latter models could not adequately account for data from human subjects, especially error RTs and the shapes of RT distributions (although Hawkins et al. found that collapsing bound models were better than the standard diffusion model for data from monkeys).

In other recent approaches in neuroscience, elements of the stimuli change rapidly over
the time course of a decision, for example, flashing lights, clicks, or odors. It has been assumed for modeling the data from these tasks that as the elements change moment to moment, drift rate changes moment to moment (Bowman et al., 2012; Brunton et al., 2013; Kira et al., 2015; Park et al., 2016). This means that if a stimulus does not contain exactly the same proportion of elements favoring one choice over the other on each trial, then this is conceptually the same as across-trial variability in drift rate. There are several issues to mention about this class of tasks. First, if the elements are presented slowly, then the task moves out of the domain of perceptual decision-making and into the domain of expanded-judgment tasks (Ratcliff, Smith, Brown, & McKoon, 2016). There is some continuity between these two classes of tasks, but it has not been possible to model the two in the same way. In the expanded-judgment domain with human subjects, there are large individual differences in how the task is performed (Smith & Vickers, 1989; see discussion in Ratcliff et al., 2016), including differential weighting of early and late information. Second, when the elements are presented very quickly, it is likely that stimulus information is integrated over time (Ratcliff & Rouder, 2000; Ratcliff et al., 2016; Smith & Ratcliff, 2009). This could provide a different explanation of the data than the models that have been implemented for these tasks (especially in dealing with RT distributions and error RTs, which can prove critical in discriminating between models). Third, it is difficult to see how the above implemented models for rapidly changing stimuli could deal with static displays (Experiments 1, 2, and 4 in this article).

More generally, there has been considerable research in neuroscience into the variability of neural responses and how that relates to overt behavior (e.g., Cohen & Maunsell, 2011; Faisal, Selen, & Wolpert, 2008; Nienborg, Cohen, & Cumming, 2012; Parker & Newsome, 1998). Repeated stimuli have been used to examine the consistency of neural responses across them and it has been found that there is a small but significant correlation between the firing rates of single neurons in MT when presented with the same stimulus (e.g., Britten, Shadlen, Newsome, & Movshon, 1992) and between pairs of neurons in MT and V5 when recorded simultaneously (e.g., Zohary, Shadlen, & Newsome, 1994). Simultaneous recording in the last stages of decision-related areas of the oculomotor system also show such correlations (Ratcliff, Hasegawa,
Hasegawa, Childers, Smith, & Segraves, 2011; see also Port & Wurtz, 2003). All these results are consistent with the earlier work of Swets and Green.

In many (but not all) implemented diffusion models of perceptual decision making in neuroscience, the models assume (perhaps implicitly) that these sources of noise collapse onto a single, internal noise component (Churchland et al., 2008; Deneve, 2012; Drugowitsch et al., 2012; Hanks et al., 2011; Kiani et al., 2014; Palmer et al., 2005; Zhang et al., 2014; but see Ditterich, 2006a, 2006b, 2008). This contrasts with many models in psychology that assume trial-to-trial variability in model components. This discrepancy possibly occurs because, in many neuroscience applications, models are not tested against the shapes and locations of RT distributions for correct or error responses and so the data that are crucial in requiring trial-to-trial variability are not addressed. For example, in an analysis of noise in decision making, Churchland et al. (2011) considered a model in which spike rates were from a Poisson process and the rate parameter varied from trial to trial. It would seem that this would be the same as trial-to-trial differences in drift rates. However, in implemented models from this domain, the assumption of trial-to-trial differences in drift rates in diffusion models is not favored in modeling in perceptual decision making (Shadlen & Kiani, 2013).

The data from this study provide the first direct behavioral evidence of systematic trial-to-trial variability in drift rate in perceptual and cognitive decision-making. We emphasize that measuring the various external and internal sources of noise requires a model that makes explicit assumptions about their contributions to performance, as the diffusion model does. Getting these assumptions right is crucial for interpreting experimental results in neuroscience and for clinical applications; incorrect assumptions may assign experimental effects to the wrong processes.
Acknowledgments

Preparation of this article was supported by NIA grant R01-AG041176. Research from this article was presented by the authors at the Psychonomic Society 57th Annual Meeting in Boston, MA, November, 2016, and at a Workshop on Sequential Sampling Models of Decision Making in Emmetten, Switzerland, May, 2016.
References


[http://doi.org/10.1037/a0034954](http://doi.org/10.1037/a0034954)


http://doi.org/10.1073/pnas.0812589106


http://doi.org/10.1037/0882-7974.18.3.415


http://doi.org/10.1037/0033-295X.108.3.550

http://doi.org/10.3758/BRM.40.1.61


http://doi.org/10.1016/j.jmp.2007.09.005


http://doi.org/10.3389/fninf.2013.00014


Figure Captions

Figure 1. A: An illustration of the diffusion model. There are three simulated paths with mean drift rate $v$, starting point $z$, and boundary separation $a$. One process hits the top boundary quickly, another hits it later, and another hits the bottom boundary in error. This also shows how the model predicts the right-skewed shapes of RT distributions: most processes hit the boundary quickly but some hit later. B: A distribution across trials of perceptual strength (or drift rate) with mean $v$ and standard deviation $\eta$. One of the red dashed and blue dotted lines is a random sample from the distribution and the other illustrates a correlated sample from the identical stimulus. C: This shows the mapping into drift rates. D: Examples of binomial random variables with probability (accuracy in the experiments) on the y-axis and the probability of agreement between a first and a repeated sample on the x-axis. The different curves represent different correlations between the two samples.

Figure 2. Example trials of the five experimental tasks. Full descriptions are in the Appendix A.

Figure 3. Quantile probability functions for the data (x’s) and model predictions (o’s and lines joining them) from the five experiments. RT quantiles are plotted against response proportions (correct responses to the right of 0.5 and errors to the left). The green central lines are the median RTs and the lines (from bottom to top) represent the 0.1, 0.3, 0.5, 0.7, and 0.9 quantile RTs. The small insert in C represents equal area rectangles drawn between quantile RTs to represent a RT distribution. Each subject was fit separately and model predictions were generated for each subject. The predictions and the data were averaged in the same way.

Figure 4. A-E show plots of accuracy against agreement between pairs of responses in the double-pass procedure. Seven values of drift rate and seven values of the standard deviation in drift rate across trials ($\eta$) were used to produce each function (shown as the small dots on the lines. Each line corresponds to one value of $\eta$ and each dot on a function corresponds to one of the seven values of drift rate. The other model parameters were the means from the fits to the data. The thick line in each plot represents the function nearest the value of $\eta$ estimated from fits.
to the model to data. The red squares are values of accuracy plotted against agreement for each condition of each experiment and the blue squares represent a replication of the numerosity discrimination experiment with the repeated stimuli randomized in order (these are the lower of each pair of squares for the numerosity discrimination experiment). F shows a bivariate density plot for response times for the first and second stimulus presentations for simulated values from the diffusion model with parameters from the numerosity discrimination task for the middle accuracy condition. The correlation between the response times is 0.107.

Figure B1 shows three examples of binomial trials with calculations of accuracy and agreement (see the Appendix B for details).

Figure C1. A: shows how the durations of the processing components add to give total response time. B: lists the model parameters. Drift rate is normally distributed across trials with SD $\eta$, starting point is uniformly distributed with range $s_z$, and nondecision time is uniformly distributed with range $s_t$. C: illustrates the mapping from response time distributions and accuracy to drift rates, boundary settings, and nondecision time in the model fitting process.
Appendix A: Detailed Experimental Methods

For all the experiments, the stimuli were displayed on the screen of a PC and responses were collected from the PC’s keyboard. Assuming a standard viewing distance of 53 cm, each pixel is 0.054 degrees on a side. Subjects were instructed to respond as quickly and accurately as possible. In all the experiments, a randomized block of trials was presented followed by an exact repetition of the block. Experiment 1 was replicated with a random order of stimuli in the repeated block (results are presented in Figure 4 and model parameters in Table 1).

Experiment 1, numerosity discrimination (Ratcliff, Thapar, & McKoon, 2010; Ratcliff, 2014): Sixteen young adults participated in the experiment for credit for an introductory psychology course. For each trial, an array of asterisks was displayed on the PC screen, and subjects were asked to determine whether the array was large or small. The positions to be filled with asterisks were chosen randomly from 100 positions laid out in a 10 x 10 array. The number of asterisks ranged from 36 to 65 with a large/small cutoff of 50. Subjects indicated whether the number of asterisks was large or small by pressing the “/” key for “large” or the “z” key for “small,” and the array remained on the screen until a response was made. Then accuracy feedback (“correct” for responses of “large” to arrays with 51 or more asterisks and for responses of “small” to arrays with 50 or fewer asterisks, or “error”) was displayed for 500 ms followed by a blank screen for 400 ms and then the next array. To discourage fast guessing, a “TOO FAST” message was displayed for 1500 ms before the blank screen for responses shorter than 280 ms. There were 18 blocks of 90 trials each. All possible numbers of asterisks were presented three times per block in random order. In the data analyses we grouped correct responses to 36-40 asterisks with correct responses to 61-65 asterisks, correct responses to 41-45 asterisks with
correct responses to 56-60 asterisks, and correct responses to 46-50 asterisks with correct responses to 51-55 asterisks (and an equivalent grouping for errors). This produced easy, medium, and difficult conditions.

**Experiment 2, letter discrimination** (Thapar et al., 2003): Seventeen young adults participated in the experiment for credit for an introductory psychology course. For each trial, one of two letters was displayed on the screen and then masked, and subjects were asked to indicate which letter was presented. The stimuli were white letters displayed in the center of the computer screen against a dark background. Letters were paired so as to be dissimilar from each other. The pairs were F/Q, P/L, W/K, B/N, T/X, and G/R. There were 18 blocks of 96 trials, with the same two letters as the response alternatives for all of the trials of a block. The pair was chosen randomly with the restriction that each was used equally often. The two letters were displayed on either side of the center of the computer screen beginning 300 ms before the first trial, and they remained on the screen throughout the block. Each trial began with a + sign in the center of the screen displayed for 500 ms, then it disappeared for 300 ms, then the target letter was displayed, followed by a variable delay (10, 20, or 30 ms) and a mask. The mask was a square outline, larger than the letter stimuli, filled with randomly placed horizontal, vertical, and diagonal lines, sampled randomly from a picture that was about 10 times larger in area than the mask and filled with randomly placed horizontal, vertical, and diagonal lines (thus the mask was different on every trial). The mask remained on the screen until a response was given. Subjects were instructed to press the “/” key if the letter on the right had been presented, and the “z” key if the letter on the left had been presented. Responses longer than 1,500 ms were followed by a “TOO SLOW” message displayed for 300 ms, and responses shorter than 250 ms were followed by a “TOO FAST” message displayed for 1,500 ms. Incorrect responses were followed by an
“ERROR” message displayed for 300 ms, and no feedback was provided for correct responses. The response alternative letters were 140 pixels each from the center, or 7.56 degrees from the center. Each stimulus/lure letter was 16x20 pixels, or 0.86 x 1.08 degrees. The line mask was 61x61 pixels, or 3.29 x 3.29 degrees.

**Experiment 3, motion discrimination** (Ratcliff & McKoon, 2008): Sixteen young adults participated in the experiment for credit for an introductory psychology course. For each trial, the screen began as a blank, black background for 300 ms. Then 24 screens of 3 interlaced coherent image sets composed of dots was displayed, with some proportion of the dots moving in the same direction (either to the left or to the right) and others switching into random positions. Subjects were asked to decide whether the dots that moved coherently were moving left or right. A series of frames were displayed on a PC screen, each frame for 16.7 ms. Five dots, each 1 by 1 pixel in size (0.054 degree square), were displayed in a circular aperture 5.4 degrees in diameter centered on the PC screen. On the first three frames, the dots were located in random positions. On the fourth through sixth frames, there was some probability that a dot moved coherently, that is, to the left or right. Specifically, for the fourth frame, dots moving coherently were probabilistically chosen from the dots that had appeared on the first frame; for the fifth frame, they were probabilistically chosen from those that had appeared on the second frame; and for the sixth frame, they were probabilistically chosen from those that had appeared on the third frame. Then the sequence continued, with the five dots on frames 7, 8, and 9 moving according to the same scheme as for frames 4, 5, and 6. Sequences like this continued for 400 ms (24 60 Hz screens), after which the stimulus disappeared to a blank screen. Dots not chosen to move coherently were positioned randomly on each frame. The probability of dots moving coherently was .10, .15, or .20. The coherently moving dots moved by 4 pixels from frame to frame, that is
13 degrees per second. With proportion and direction (left or right), there were 6 conditions in the experiment. There were 18 blocks in each session with 96 trials per block. For each block, each of the 6 conditions was tested 16 times, in random order. If a response was correct or incorrect, the word “correct” or “ERROR” was displayed for 300 ms followed by a 300 ms blank screen. For responses shorter than 280 ms, a “TOO FAST” message was displayed for 1,500 ms before the blank screen. For responses slower than 1500 ms, a “TOO SLOW” message was displayed for 300 ms before the blank screen. The dots appeared anywhere within a 100 pixel, or 5.40 degree disk at the center of the screen. Each dot was 1x1 pixel, or 0.054 x 0.054 degrees. A coherent move was 4 pixels, or 0.22 degrees to the left or right.

**Experiment 4, static brightness discrimination** (Ratcliff & Smith, 2010; Ratcliff, Thapar, & McKoon, 2003; Ratcliff, 2002, 2014): Seventeen young adults participated in the experiment for credit for an introductory psychology course. For each trial, a square of black and white pixels was displayed on a gray background then masked, and subjects were asked to decide if the square was “bright” or “dark.” The brightness of each square was manipulated by varying the proportion of pixels that were white. At the beginning of each block of trials, a 320 x 200 pixel, 50% grayscale background was displayed, and remained on the screen throughout the block. Each trial began with a + sign fixation point displayed on the gray background for 250 ms. Then the + sign disappeared and a 64 x 64 pixel stimulus was immediately displayed in the center of the background for 100 ms. Four brightness levels were used: .43, .46, .54, or .57 white pixels. Then four checkerboard mask patterns, each 64 x 64 pixels, were displayed in the following order: a checkerboard with 2 x 2 black and white squares, a checkerboard the same as the first but with the black and white squares reverse, a checkerboard with 3 x 3 black and white squares, and then its reverse. The checkerboards were designed to mask both smaller and larger
random features of a stimulus that might have remained visible through only one or two of the masks. The last checkerboard remained on the screen until a response was given. Subjects were instructed to press the “/” key if they judged the stimulus to be “bright” and the “z” key for “dark.” If a response was correct, the display reverted to the gray background for 500 ms and then the next stimulus was displayed. If the response was incorrect, “ERROR” was displayed for 300 ms before the gray background. If a response was slower than 2000 ms, a “TOO SLOW” message was displayed for 500 ms. To discourage fast guessing, a “TOO FAST” message was displayed for 1,500 ms if the response was shorter than 250 ms, just before the gray background. There were 18 blocks of 96 trials each, with all four brightness levels tested an equal number of times in each block, in random order.

Experiment 5, dynamic brightness discrimination (Ratcliff & Smith, 2010): Nineteen young adults participated in the experiment for credit for an introductory psychology course. Subjects were presented with a dynamic, homogeneous, random 60 x 60 array of black and white pixels, with the proportion of black to white pixels the same for each frame with different random samples of pixels presented in consecutive frames, and were asked to judge whether the array was “bright” or “dark.” For each trial, a + sign fixation point was displayed for 500 ms in the center of a 320 x 200 pixel 50% grayscale background. The dynamic stimulus was then immediately displayed, and continued until a response was given. Each stimulus consisted of a sequence of frames presented at a frame rate of 60 Hz, with a new set of contrast-reversed pixels randomly chosen on each frame. Four brightness levels were used: .46, .48, .52, or .54 white pixels. Subjects were instructed to press the “/” key if they judged the stimulus to be “bright” (more white pixels than black pixels), and the “z” key if they judged the stimulus to be “dark.” If the response was incorrect, “ERROR” was displayed for 300 ms. No feedback was given for
correct responses. If a response was longer than 2000 ms, a “TOO SLOW” message was displayed for 500 ms. To discourage fast guessing, a “TOO FAST” message was displayed for 1,500 ms just before the gray background if the response was shorter than 250 ms. There were 20 blocks of 96 trials each, with each brightness level presented 24 times per block.

For all the experiments, data from the first and second blocks and the first response in each block were discarded from the analysis. Responses with RTs less than 300 ms and above 2000 ms were eliminated from analyses.
Appendix B: Examples of Accuracy and Agreement for Binary Responses

Figure B1 shows three examples of agreement and accuracy. Block 1 in each are responses on the first presentation of an item and Block 2 are responses for the second. For the first two examples, responses on the two presentations are independent and the binomial probabilities are 0.5 and 0.9, respectively. For the third example, the binomial probability is 0.5 and there is a correlation between the two blocks of 0.7. These examples illustrate how the functions in Figure 1E come about. The first two examples show accuracy values of 0.5 and 0.9 with 0 correlation which produce agreement values of 0.5 and 0.9. The third example shows accuracy values of 0.5 with 0.7 correlation which produces an agreement value of 0.9.
Appendix C: Fitting the Diffusion Model

The standard diffusion model is designed to explain the cognitive processes involved in making simple two-choice decisions, decisions that take place in under a second or two, and to explain all the data for the decisions: accuracy and the full distributions of RTs (their shapes and locations) for correct responses and errors. Decisions are made by a noisy process that accumulates information over time from a starting point toward one of the two boundaries. The rate of accumulation of information is called drift rate and it is determined by the quality of the information extracted from the stimulus in perceptual tasks and the quality of match between the test item and memory in lexical decision and memory tasks. In Figure 1A, the boundaries are \( a \) and \( 0 \), the starting point is \( z \), and the drift rate is shown for a condition for which the correct decision is at the top boundary. Processes outside the decision process such as stimulus encoding, mapping the stimulus representation to a decision-relevant representation, and response execution are combined into one component of the model, labeled “nondecision” time, with mean \( T_{cr} \). Total RT is the sum of the time to reach a boundary and the nondecision time (Figure C1A). The noise in the accumulation of information, “within-trial” (internal) variability (Gaussian distributed noise), results in decision processes with the same mean drift rate terminating at different times (producing RT distributions) and sometimes at the wrong boundary (producing errors).

The values of the components of processing are assumed to vary from trial to trial, under the assumption that subjects cannot accurately set the same parameter values from one trial to another (e.g., Laming, 1968; Ratcliff, 1978). Across-trial variability in drift rate is normally distributed with SD \( \eta \), across-trial variability in starting point is uniformly distributed with range
s_z, and across-trial variability in the nondecision component is uniformly distributed with range s_t. Across-trial variability in drift rate and starting point allow the model to fit the relative speeds of correct and error responses (Ratcliff, Van Zandt, & McKoon, 1999; Ratcliff & McKoon, 2008). In signal detection theory, which deals only with accuracy, all sources of across-trial variability are collapsed into one parameter, the variability in information across trials. In contrast, with the diffusion model, the separate sources of across-trial variability are identified. Figure C1B shows the model parameters.

Boundary settings, nondecision time, starting point, the drift rates for each condition, and the across-trial variabilities in drift rate, nondecision time, and starting point are all identifiable. When data are simulated from the model (with numbers of observations approximately those that would be obtained in real experiments) and the model is fit to the data, the parameters that were used to generate the data are well recovered (Ratcliff & Tuerlinckx, 2002). The success of parameter identifiability comes in part from the tight constraint that the model account for the full distributions of RTs for correct and error responses (Ratcliff, 2002).

To fit the diffusion model to the data, the RT distributions were represented by 5 quantiles, the .1, .3, .5, .7, and .9 quantiles. The quantiles and the response proportions were entered into the minimization routine and the diffusion model was used to generate the predicted cumulative probability of a response occurring by that quantile RT. Subtracting the cumulative probabilities for each successive quantile from the next higher quantile gives the proportion of responses between adjacent quantiles. For a G-square computation, these are the expected proportions, to be compared to the observed proportions of responses between the quantiles (i.e., the proportions between 0, .1, .3, .5, .7, .9, and 1.0, which are .1, .2, .2, .2, .2, and .1). The
proportions for the observed ($p_o$) and expected ($p_e$) frequencies and summing over $N * p_o * \log(p_o/p_e)$ for all conditions gives a single G-square value to be minimized. This is accomplished using a general SIMPLEX minimization routine. The parameter values for the model are adjusted by SIMPLEX until the minimum G-square value is obtained. The number of degrees of freedom in the data are the 12 proportions between the quantiles and extremes (6 each for correct and error responses) minus 1 (because the sum must equal 1) multiplied by the number of conditions in the data. The model was fit to the data for each subject individually the same way as fitting the chi-square method described by (Ratcliff & Tuerlinckx, 2002; see also Ratcliff & Childers, 2015). In fitting the model to data, we collapsed correct responses to one choice with correct responses to the other choice (and the same for corresponding errors) as has been done in other studies (see Ratcliff, 2014). This can be done because responses to one choice are symmetric with responses for the other choice and so the starting point of the diffusion process is set to be midway between the boundaries. In fitting the model, the values of all the parameters, including the variability parameters, are estimated simultaneously, for all the data from all the conditions of an experiment (see Figure C1C).
Table 1: Diffusion Model Parameters from Fits to Data

<table>
<thead>
<tr>
<th>Discrimination task</th>
<th>a</th>
<th>Ter</th>
<th>η</th>
<th>s_e</th>
<th>s_t</th>
<th>v_1</th>
<th>v_2</th>
<th>v_3</th>
<th>G^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numerosity</td>
<td>0.129</td>
<td>0.411</td>
<td>0.124</td>
<td>0.068</td>
<td>0.197</td>
<td>0.074</td>
<td>0.225</td>
<td>0.344</td>
<td>43.0</td>
</tr>
<tr>
<td>Numerosity (random)</td>
<td>0.127</td>
<td>0.396</td>
<td>0.153</td>
<td>0.073</td>
<td>0.178</td>
<td>0.072</td>
<td>0.220</td>
<td>0.348</td>
<td>39.4</td>
</tr>
<tr>
<td>Letter</td>
<td>0.111</td>
<td>0.381</td>
<td>0.209</td>
<td>0.032</td>
<td>0.160</td>
<td>0.104</td>
<td>0.260</td>
<td>0.364</td>
<td>58.2</td>
</tr>
<tr>
<td>Motion</td>
<td>0.098</td>
<td>0.438</td>
<td>0.156</td>
<td>0.062</td>
<td>0.238</td>
<td>0.098</td>
<td>0.139</td>
<td>0.191</td>
<td>43.0</td>
</tr>
<tr>
<td>Static brightness</td>
<td>0.104</td>
<td>0.431</td>
<td>0.164</td>
<td>0.044</td>
<td>0.208</td>
<td>0.173</td>
<td>0.274</td>
<td></td>
<td>31.0</td>
</tr>
<tr>
<td>Dynamic brightness</td>
<td>0.113</td>
<td>0.423</td>
<td>0.157</td>
<td>0.070</td>
<td>0.230</td>
<td>0.135</td>
<td>0.246</td>
<td></td>
<td>31.4</td>
</tr>
</tbody>
</table>

The parameters were: Boundary separation $a$ (starting point $z = a/2$), mean nondecision component of response time, $T_{er}$, SD in drift across trials $\eta$, range of the distribution of starting point $s_e$ and range of the distribution of nondecision times, $s_t$. 95% critical values of chi-squares ($G^2$ is asymptotically distributed $\chi^2$ square) are 37.6 for 25 degrees of freedom for the numerosity, letter, and motion discrimination tasks and 25.0 for 15 degrees of freedom for the brightness discrimination tasks. Values of the mean $\chi^2$ between 1 and 2 times the critical value are representative of adequate fits (Ratcliff and Childers, 2015). For drift rates, $v_1$ represents the most difficult condition, $v_2$ easier, and $v_3$ the easiest condition.
Figure 1

A. Correct RT distribution

B. Error RT distribution

C. Drift Rate

D. Accuracy vs. Agreement

0 0.2 0.4 0.6 0.8 1.0

0.5 0.6 0.7 0.8 0.9 1.0

Correlation

Agreement

Correct response

Drift Rate v

Error response

0 v
Figure 4

A: Numerosity Discrimination
B: Letter Discrimination
C: Motion Discrimination
D: Static Brightness Discrimination
E: Dynamic Brightness Discrimination
F: Heatmap showing RT for first and second presentation
Appendix B Figure 1

Binomial probability = 0.5

Block 1: 0 1 1 1 0 0 1 0 0 1 1 1 1 0 0 1 0 1 0 0
Block 2: 1 0 0 1 0 1 0 1 0 1 0 1 0 0 1 1 0 1 1 0
Agreement = 0+0+0+1+1+...

....0+1 Agreement Prob = 10/20 = 0.5

Binomial probability = 0.9

Block 1: 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 0 1 1
Block 2: 1 1 0 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1
Agreement = 1+1+0+1+1+...

....1+1 Agreement Prob = 18/20 = 0.9

Binomial probability = 0.5, but a correlation of 0.7

Block 1: 1 0 1 1 0 0 1 1 1 0 0 1 0 1 1 0 1 0 0 0
Block 2: 1 0 1 0 0 0 1 1 1 1 0 1 0 1 1 0 1 0 0 0
Agreement = 1+1+1+0+1+...

....1+1 Agreement Prob = 18/20 = 0.9
Appendix C Figure 1

A

Total RT = u + d + w

\[ \text{d} \quad \text{decision} \]

\[ u \quad \text{encoding, transformation, etc.} \]

\[ w \quad \text{response output} \]

Time

Nondecision components of RT = u + w

mean = T_{er}

B

Parameters of the Diffusion Model

1. \( a \) = Boundary separation
2. \( z \) = Starting point
3. \( T_{er} \) = Mean encoding and response time
4. \( v \) = Mean drift rate (one for each condition)
5. \( \eta \) = Standard deviation in mean drift rate from trial to trial (normal distribution)
6. \( s_z \) = Range in starting point \( z \) from trial to trial (uniform distribution)
7. \( s_t \) = Range in \( T_{er} \) (uniform distribution)

C

Data (one set for each condition)

<table>
<thead>
<tr>
<th>Condition 1</th>
<th>Condition 2</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>Accuracy</td>
<td></td>
</tr>
<tr>
<td>Correct RT distribution</td>
<td>Correct RT distribution</td>
<td></td>
</tr>
<tr>
<td>Error RT distribution</td>
<td>Error RT distribution</td>
<td></td>
</tr>
</tbody>
</table>

Model components

- Drift rates \( v_1, v_2, \ldots \)
  - (evidence from perception or memory; one for each condition)
- Boundary separation \( a \)
  - (amount of evidence needed for a decision)
  - Starting point \( z \)
- Nondecision time \( T_{er} \)
  - (duration of encoding, extracting decision related information, motor process, etc.)

Across trial variability in model components:
1. \( \eta \) drift rate,
2. \( s_z \) starting point,
3. \( s_t \) nondecision time.