Two-Stage Dynamic Signal Detection:
A Theory of Choice, Decision Time, and Confidence

Timothy J. Pleskac
Michigan State University

Jerome R. Busemeyer
Indiana University

PLEASE DO NOT CITE WITHOUT PERMISSION
Dr. Timothy J. Pleskac
Dept. of Psychology
Michigan State University
East Lansing, MI, 48823
517.353.8918
	tim.pleskac@gmail.com

January 21, 2010
Abstract
The three most-often-used performance measures in the cognitive and decision sciences are choice, response or decision time, and confidence. We develop a random walk/diffusion theory – the Two-stage Dynamic Signal Detection (2DSD) theory – that accounts for all three measures using a common underlying process. The model uses a drift diffusion process to account for choice and decision time. To estimate confidence, we assume that evidence continues to accumulate after the choice. Judges then interrupt the process to categorize the accumulated evidence into a confidence rating. The model explains all known interrelationships between the three indices of performance. Furthermore, the model also accounts for the distributions of each variable in both a perceptual and general knowledge task. The dynamic nature of the model also reveals the moderating effects of time pressure on the accuracy of choice and confidence. Finally, the model specifies the optimal solution for giving the fastest choice and confidence rating for a given level of choice and confidence accuracy. Judges are found to act in a manner consistent with the optimal solution when making confidence judgments.

Keywords: Confidence, Diffusion Model, Subjective Probability, Optimal Solution, Time Pressure
Confidence has long been a measure of cognitive performance used to chart the inner workings of the mind. For example, in psychophysics confidence was originally thought to be a window onto Fechner’s perceived interval of uncertainty (Pierce, 1877). At the higher levels of cognition, confidence ratings about recognition are used to test and compare different theories of memory (e.g., Ratcliff, Gronlund, & Sheu, 1992; Squire, Wixted, & Clark, 2007; Yonelinas, 1994). Confidence has also been used in the decision sciences to map the correspondence between a person’s internal beliefs and reality, whether it be the accuracy of meteorologists’ forecasts (Murphy & Winkler, 1977), the accuracy of students predicting the proportion of correct and incorrect responses on a test (Lichtenstein, Fischhoff, & Phillips, 1982), or the accuracy of a local sports fan predicting the outcome of games (Yates & Curley, 1985).

This common reliance on confidence implies that the cognitive and decision sciences each have a vested interest in understanding confidence. Yet, on closer inspection our understanding of confidence is limited. For instance, an implicit assumption amongst most psychological theories is that observed choices, decision times, and confidence ratings, tap the same latent process. Most successful cognitive models, however, only account for two of these three primary measures of performance. For example, signal detection models assume confidence ratings differ from choice only in terms of the “response set available to the observer” (Macmillan & Creelman, 2005, p. 52). Signal detection theory, however, is silent in terms of decision time. As a result, random walk/diffusion theory was introduced as an explanation of both choices and decision times (Laming, 1968; Link & Heath, 1975; Ratcliff, 1978; Stone, 1960). A great limitation of random walk/diffusion theory, however, is its inability to account for confidence ratings (Van Zandt, 2000b; Van Zandt & Maldonado-Molina, 2004; Vickers, 1979). So this leaves us with a challenge – is it possible to extend the random walk/diffusion class of models to account for confidence? In this article we address this challenge by developing a ‘dynamic signal detection theory’ that combines the strengths of a signal detection model of confidence with the power of random walk/diffusion theory to model choice and decision time.

Such a dynamic understanding of confidence has a number of applications. In this article, we use our dynamic understanding of confidence to address an important question involving confidence: what is the effect of time and time pressure on the accuracy of our confidence
ratings? To address this question we use methods developed in the decision sciences where confidence ratings are treated as subjective probabilities (Adams, 1957; Adams & Adams, 1961; Lichtenstein, et al., 1982; Tversky & Kahneman, 1974). The accuracy of subjective probabilities has been well studied in the decision sciences (for reviews see Arkes, 2001; Griffin & Brenner, 2004; Koehler, Brenner, & Griffin, 2002; McClelland & Bolger, 1994). Yet, little is known as to how or why the accuracy of subjective probability estimates might change under time pressure and more generally how judges balance time and accuracy in producing not only choice, but also confidence ratings.

To build this dynamic understanding of confidence, we focus on situations where judges face a standard detection task where they are shown a stimulus and are asked to choose between two alternatives (A or B). Judges are uncertain about the correct response. For example, an eyewitness may have to decide if a face in a photograph was present at the crime scene, a military analyst may have to decide whether a particular target is a threat, or a test-taker may have to decide if a statement is true. After making a choice, judges express their confidence in their choice. According to our theory, judges complete this task of making a choice and entering a confidence rating in two stages (Figure 1).

In the first stage (left the vertical line located at $t_D$ in Figure 1), judges make a choice based on a sequential sampling process best described by random walk/diffusion theory (Laming, 1968; Link & Heath, 1974; Ratcliff, 1978; Stone, 1960). During this stage, judges begin to sequentially accumulate evidence favoring one alternative over the other. Typically, the evidence has some direction or drift where it favors one alternative over the other. The path of the accumulated evidence is shown as a jagged line in Figure 1. If the evidence drifts upward it favors response alternative A and if it drifts downward it favors B. The information serving as evidence can come from one of many different sources including current sensory inputs and/or memory stores. The jagged line in Figure 1 also illustrates that the sampled evidence at each time step is subject to random fluctuations. This assumption also characterizes the difference between random walk and diffusion models. In random walk models the evidence is sampled in discrete time intervals; while in diffusion models, the evidence is sampled continuously in time.

When judges reach a preset level of evidence favoring one alternative over the other, they stop collecting evidence and make a choice accordingly. Thus, this process models an optional stopping choice task where observers control their own sampling by choosing when they are
ready to make a choice. An alternative task is an *interrogation choice task* where an external event (e.g., experimenter) interrupts judges at different sample sizes and asks them for a choice. This interrogation choice task can also be modeled with random walk/diffusion theory (see Ratcliff, 1978; Roe, Busemeyer, & Townsend, 2001).

Returning to the optional stopping model, the horizontal lines labeled with $\theta_A$ and $\theta_B$ in Figure 1 depict the preset level of evidence or thresholds for the two different choice alternatives. These thresholds are typically *absorbing boundaries* where once the evidence reaches the threshold the accumulation process ends (Cox & Miller, 1965). If we make a final assumption that each sampled piece of evidence takes a fixed amount of time, then random walk/diffusion theory models *decision times* as the time it takes a judge to reach $\theta_A$ or $-\theta_B$ (the first passage time).

In summary, random walk/diffusion theory models both choice and decision times as a compromise between the (a) the *quality* of the accumulated evidence as indexed by the drift or drift rate of the evidence; and (b) the *quantity* of the accumulated evidence as indexed by the choice thresholds (Ratcliff & Smith, 2004). Models based on these three plausible assumptions have been used to model choices and decision times in sensory detection (Smith, 1995), perceptual discrimination (Link & Heath, 1975; Ratcliff & Rouder, 1998), memory recognition (Ratcliff, 1978), categorization (Ashby, 2000; Nosofsky & Palmeri, 1997), risky decision making (Busemeyer & Townsend, 1993; J. G. Johnson & Busemeyer, 2005) and multi-attribute, multi-alternative decisions (Diederich, 1997; Roe, Busemeyer, & Townsend, 2001) as well as other types of response tasks like the Go/No-Go task (Gomez, Perea, & Ratcliff, 2007).

Often, though, an alternative measure of cognitive performance in the form of confidence is collected. Confidence is easily obtained with a simple adjustment in the empirical procedure: after judges make a choice ask them to rate their confidence that their choice was correct. Indeed this simple empirical adjustment to collect confidence ratings has allowed psychologists to gain insight in a variety of areas including estimating empirical thresholds in psychophysics (Link, 1992; Pierce & Jastrow, 1884), examining the role of consciousness and our own awareness of our cognitions (Nelson, 1996, 1997; Nelson & Narens, 1990), or simply as a method to collect data in an efficient manner for a variety of tasks (Egan, 1958; Green & Swets, 1966; Ratcliff et al., 1992).
In this paper, we focus on the confidence with choice task. This focus allows us to explicitly model how choice and decision time are related to confidence. Moreover by simultaneously modeling choice, decision time, and confidence, we investigate the degree to which the same dynamic process can account for all three measures of cognitive performance. Of course confidence could be collected in a slightly different manner, with a no-choice task where experimenters simply ask judges to rate their confidence that a particular item (e.g., A) is correct. In the discussion, we address how our model might be adapted to account for these tasks as well (see also Ratcliff & Starns, 2009).

To model confidence we deviate from the often made assumption that confidence in choice is derived from the same accumulated evidence that led to a choice (though see Busey, Tunnicliff, Loftus, & Loftus, 2000). This assumption is true for signal detection models of confidence (e.g., Macmillan & Creelman, 2005; Suantek, Bolger, & Ferrell, 1996; Budescu, Wallsten, & Au, 1997) as well as a majority of sequential sampling models of confidence (e.g., Heath, 1984; Link, 2003; Vickers, 1979; 2001; Van Zandt, 2000b; Merckle & Van Zandt, 2006). We take a different course. We propose, as Baranski and Petrusic (1998) first suggested, that there is post-decisional processing of confidence judgments. Our hypothesis is that this post-decision processing takes the form of continued evidence accumulation in terms of the two possible responses (see also Van Zandt & Maldonado-Molina, 2004). In other words, we remove the assumption that choice thresholds are absorbing boundaries. Instead the threshold indicates a choice to be made, and then the confidence is derived from a second stage of evidence accumulation that builds on the evidence accumulated during the choice stage.

We call this general theory of choice and confidence unfolding over two separate stages two-stage dynamic signal detection theory (2DSD). The right half of Figure 1 (right of the vertical line placed at \(t_D\)) depicts one realization of the second stage of evidence accumulation where the stopping rule for confidence is modeled with an interrogation-type stopping rule. We call this model the 2DSD interrogation model where after making a choice, judges then interrupt the second stage of the evidence accumulation after a fixed amount of time \(\tau\) or inter-judgment time. Judges then use the state of evidence to select a confidence rating accordingly.

An alternative stopping rule for the second stage is that judges use a different stopping rule akin to an optional stopping rule where they lay out markers along the evidence state space representing the different confidence ratings. The markers operate so that each time the
accumulated evidence passes one of these markers there is a probability that the judge exits and gives the corresponding confidence rating. We call this model the 2DSD optional stopping model. The advantage of this version of 2DSD is that it simultaneously models choice, decision time, confidence, and inter-judgment time distributions.

The 2DSD interrogation model, however, is conceptually and formally easier to work with. Thus, in this article, we will initially rely on the interrogation model to investigate the implications of the 2DSD framework in general. Later we will investigate the degree to which a 2DSD optional stopping model can account for the data, in particular inter-judgment times. The more important psychological aspect of both models is that in order to understand choice, decision time, and confidence one has to account for the evidence accumulated in the second stage of processing.

Our development of 2DSD is structured as follows. As a first step, we review past attempts to model confidence within random walk/diffusion theory and examine the empirical phenomena they explain and fail to explain. This model comparison identifies the weaknesses of past attempts to model confidence and also solidifies the critical empirical phenomena or hurdles summarized in Table 1 that any cognitive model of confidence must address. Many of these hurdles were first marshaled out by Vickers (1979; see also Vickers, 2001). Furthermore, we test new predictions made by 2DSD with a new study that examines how confidence changes under different levels of time pressure in two different decision making tasks. We also use the study to help understand how the time course of confidence judgments affects the correspondence between reality and our internal subjective beliefs about events occurring. Finally, we evaluate how well the 2DSD optional stopping model can give a more precise account of the dynamic process of both choice and confidence.

Dynamic Signal Detection Theory

A common model of decision making is Green and Swets’ (1966) signal detection theory. Random walk/diffusion theory can in fact be understood as a logical extension of signal detection theory (Busemeyer & Diederich, 2009; Link & Heath, 1975; Pike, 1973; Ratcliff & Rouder, 2000; Smith, 2000; Wagenmakers, van der Maas, & Grasman, 2007). In particular, the decision process in signal detection theory can be understood as using a fixed sample size of evidence to elicit a decision (akin to the stopping rule for interrogation choice tasks). Random walk/diffusion theory, in comparison, drops this assumption. Because of this logical connection
to signal detection theory we have adopted the name dynamic signal detection theory to describe
this more general framework of models. Formally, dynamic signal detection (DSD) theory
assumes that each time interval $\Delta t$ that passes after stimulus $S_i$ ($i = A, B$) is presented, judges
consider a piece of information $y(t)$. After a time length of $t = n(\Delta t)$ judges will have generated a
set of $n$ pieces of information drawn from some distribution $f_i[y(t)]$ characterizing the stimulus.
Judges are assumed to transform each piece of information into evidence favoring one alternative
over the other, $x(t) = h[y(t)]$. Because $y(t)$ is independent and identically distributed, $x(t)$ is also
independent and identically distributed. Each new sampled piece of evidence $x(t + \Delta t)$ updates
the total state of evidence $L(t)$ so that at time $t + \Delta t$ the new total state of the evidence is,

$$L(t + \Delta t) = L(t) + x(t + \Delta t).$$

In a DSD model, when a stimulus is present, the observed information – again coming
from either sensory inputs or memory retrieval – is transformed into evidence, $x(t) = h[y(t)]$. This transformation allows DSD to encapsulate different processing assumptions. This includes
the possibility that the evidence is (a) some function of the likelihood of the information in
respect to the different response alternatives (Edwards, 1965; Laming, 1968; Stone, 1960), (b)
based on a comparison between the sampled information and a mental standard (Link & Heath,
1975), (c) a measure of strength based on a match between a memory probe and memory traces
stored in long term memory (Ratcliff, 1978), or (d) even the difference in spike rate from two
neurons (or two pools of neurons) (Gold & Shadlen, 2001; 2002). Regardless of the specific
cognitive/neurological underpinnings, according to the theory when stimulus $S_A$ is presented the
evidence is independent and identically distributed with a mean equal to $E[x(t)] = \delta \Delta t$ (the
mean drift rate) and variance equal to $V[x(t)] = \sigma^2 \Delta t$ (the diffusion rate). When stimulus $S_B$
is presented the mean is equal to $E[x(t)] = -\delta \Delta t$ and variance $V[x(t)] = \sigma^2 \Delta t$. Using Equation 1 the change in evidence can be written as a stochastic linear difference equation:

$$dL(t) = L(t + \Delta t) - L(t) = x(t + \Delta t) = \delta \Delta t + \sqrt{\Delta t} \varepsilon(t + \Delta t)$$

Where $\varepsilon(t)$ is a white noise process with a mean of 0 and variance $\sigma^2$. A standard Wiener
diffusion model is a model with evidence accruing continuously over time, which is derived
when the time step $\Delta t$ approaches zero so that the discrete process converges to a continuous
time process (Diederich & Busemeyer, 2003; Cox & Miller, 1965; Smith, 2000). A consequence
of $\Delta t$ approaching zero is that via the central limit theorem the location of the evidence accumulation process becomes normally distributed, $L(t) \sim N[\mu(t), \sigma^2(t)]$.

The DSD model also has the property that, if the choice thresholds are removed the mean evidence state increases linearly with time

$$E[L(t)] = \mu(t) = n \cdot \Delta t \cdot \delta = t \cdot \delta$$

(3)

and so does the variance

$$V[L(t)] = \sigma^2(t) = n \cdot \Delta t \cdot \sigma^2 = t \cdot \sigma^2$$

(4)

(see Cox & Miller, 1965). Thus, a measure of standardized accuracy analogous to $d'$ in signal detection theory is

$$d'(t) = 2\mu(t)/\sigma(t) = 2(\delta/\sigma)\sqrt{t} = d\sqrt{t}$$

(5)

In words, Equation 5 states that accuracy grows as a square root of time so that the longer people take to process the stimuli the more accurate they become. Equation 5 displays the limiting factor of signal detection theory. Namely that accuracy and processing time are confounded in tasks where processing times systematically change across trials. As a result the rate of evidence accumulation $d$ is a better measure of the quality of the evidence indexing the judges’ ability to discriminate between the two types of stimuli per unit of processing time. Later, we will use these properties of an increase in mean, variance, and discriminability to test 2DSD.

To make a choice, evidence is accumulated to either the upper ($\theta_A$) or lower ($-\theta_B$) thresholds. Alternative $A$ is chosen once the accumulated evidence crosses its respective thresholds, $L(t) > \theta_A$. Alternative B is chosen when the process exceeds the lower threshold, $L(t) < -\theta_B$. The time it takes for the evidence to reach either threshold or the first passage time is the predicted decision time, $t_D$. This first passage account of decision times accounts for the positive skew of response time distributions (cf., Bogacz et al., 2006; Ratcliff & Smith, 2004). The model accounts for biases judges might have toward a choice alternative with a parameter $z$, the state of evidence at time point 0, $z = L(0)$. In this framework, if $z = 0$ observers are unbiased, if $z < 0$ then observers are biased to choose alternative B, and if $z > 0$ then they are biased to respond hypothesis alternative $A$.

The utility of DSD models rests not only with the fact that it models choice and decision time, but it also models the often observed relationship between these two variables known as the speed-accuracy tradeoff (D. M. Johnson, 1939; Pachella, 1974; Schouten & Bekker, 1967;
Wickelgren, 1977). The speed-accuracy tradeoff captures the idea that in the standard detection task, the decision time is negatively related with their error rate. That is, the faster judges proceed in making a choice the more errors they make. This negative relationship produces the speed-accuracy tradeoff where judges trade accuracy for speed (Luce, 1986). The speed-accuracy tradeoff is hurdle 1 (Table 1). Any model of choice, decision time, and confidence, must account for the speed-accuracy tradeoff so often observed in decision making.

The speed-accuracy tradeoff is modeled with the threshold parameter \( \theta_i \). Increasing the magnitude of \( \theta_i \) will increase the amount evidence needed to reach a choice. This reduces the impact that random fluctuations in evidence will have on choice and as a result increase choice accuracy. Larger \( \theta_i \), however, also imply more time will be needed before sufficient evidence is collected. In comparison, decreasing the thresholds \( \theta_i \) lead to faster responses but also more errors. Assuming judges want to minimize decision times and error rates, this ability of the threshold to control decision times and error rates also implies that for a given error rate the model yields the fastest decision. In other words, the DSD model is optimal in that it delivers the fastest decision for a given level of accuracy (Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006; Edwards, 1965; Wald & Wolfowitz, 1948).

This optimality of DSD was first identified within statistics in what has been called the *sequential probability ratio test* (SPRT). The SPRT was developed to understand problems of optional stopping and sequential decision making (Barnard, 1946; Wald, 1947; Wald & Wolfowitz, 1948). Later the SPRT framework was adapted as a descriptive model of sequential sampling decisions with the goal of understanding decision times (Edwards, 1965; Laming, 1968; Stone, 1960). The SPRT model is also the first random walk/diffusion model of confidence ratings and illustrates why this class of models in general have been dismissed as a possible way to model confidence ratings (Vickers, 1979; Van Zandt, 2000b).

**Sequential Probability Ratio Tests (SPRT)**

Strictly speaking the SPRT model is a random walk model where evidence is sampled at discrete time intervals. The SPRT model assumes, that at each time step, judges compare the conditional probabilities of their information \( y(t + \Delta t) \), for either of the two hypotheses \( H_j \) (\( j = A \) or \( B \)) or choice alternatives (Bogacz et al., 2006; Edwards, 1965; Laming, 1968; Stone, 1960). Taking the natural log of the ratio of these two likelihoods forms the basis of the accumulating evidence in the SPRT model,
\[ x(t) = h[y(t)] = \ln \left[ \frac{f_A(y[t])}{f_B(y[t])} \right] \] (6)

If \( x(t) > 0 \) then this is evidence that \( H_A \) is more likely and if \( x(t) < 0 \) then \( H_B \) is more likely. Thus, the total state of evidence is tantamount to accumulating the log likelihood ratios over time

\[ L(t + \Delta t) = L(t) + \ln \left[ \frac{f_A(y[t + \Delta t])}{f_B(y[t + \Delta t])} \right] \] (7)

This accumulation accords with the log odds form of Bayes’ rule,

\[ \ln \left[ \frac{p(H_A | D)}{p(H_B | D)} \right] = \sum_t \ln \left[ \frac{f_A(y[t])}{f_B(y[t])} \right] + \ln \left[ \frac{p(H_A)}{p(H_B)} \right]. \] (8)

Judges continue to collect information so long as \(-\theta_B < L(t) < \theta_A\). This formulation is optimal in that across all fixed or variable sample decision methods, the SPRT guarantees for a given set of conditions the fastest decision time and for a given error rate (Bogacz et al., 2006; Edwards, 1965; Wald, 1947). Therefore, reaching a choice threshold (either \( \theta_A \) or \( \theta_B \)) is equivalent to reaching a fixed level of posterior odds that are just large enough in magnitudes for observers to make a choice.

The SPRT diffusion model has some empirical validity. Stone (1960) and Edwards (1965) used the SPRT diffusion model to describe human choice and decision times, although it fails to explain differences between mean correct and incorrect decision times (Link & Heath, 1975; Vickers, 1979). Gold and Shadlen (2001; 2002) have also worked to connect the SPRT model to decision making at the level of neuronal firing. In terms of confidence, the model naturally predicts confidence if we assume judges transform their final internal posterior log odds (Equation 8) with a logistic transform to a subjective probability of being correct. However, this rule (or any related monotonic transformation of the final log odds into confidence) implies that confidence is completely determined by the threshold values (\( \theta_A \) or \( \theta_B \)) or the quantity of evidence needed to make a choice. This predicted relationship between confidence and choice thresholds is problematic because when choice thresholds remain fixed across trials then this would imply that “all judgments (for a particular choice alternative) should be made with an equal degree of confidence” (Vickers, 1979, p. 175).

This prediction is clearly false and is negated by a large body of empirical evidence showing two things. The first of which is that confidence changes with the discriminability of the
stimuli. That is, confidence in any particular choice alternative is related to objective measures of
difficulty for discriminating between stimuli (Ascher, 1974; Baranski & Petrusic, 1998;
Festinger, 1943; Garrett, 1922; D. M. Johnson, 1939; Pierce & Jastrow, 1884; Pierrel & Murray,
1963; Vickers, 1979). This positive relationship between stimulus discriminability and observed
certainty is hurdle number 2 in Table 1.

A further difficulty for the SPRT account of confidence is that the resolution of confidence is usually good. That is, judges’ confidence ratings discriminate between correct and incorrect responses (e.g., Ariely et al., 2000; Baranski & Petrusic, 1998; Dougherty, 2001; Henmon, 1911; Garrett, 1922; D. M. Johnson, 1939; Nelson & Narens, 1990; Vickers, 1979). This resolution even remains when stimulus difficulty is held constant (Baranski & Petrusic, 1998; Henmon, 1911). In particular, judges typically have greater confidence in correct choices than in incorrect choices (hurdle 3). The SPRT model, however, predicts equal confidence for correct and incorrect choices.4

In sum, the failures of the SPRT model reveal that any model of confidence must account for the monotonic relationship between confidence and an objective measure of stimulus difficulty as well as the relationship between accuracy and confidence. These two relationships serve as hurdles 2 and 3 for models of confidence (Table 1). Furthermore, the SPRT model demonstrates that the quantity of accumulated evidence as indexed by the choice threshold ($\theta$) is not sufficient to account for confidence and thus serves as an important clue in the construction of a random walk/diffusion model of confidence. An alternative model of confidence treats confidence as some function of both the quality ($\delta$) and quantity of evidence collected ($\theta$). In fact, this hypothesis has its roots in C. S. Pierce’s (1877) model of confidence – perhaps one of the very first formal hypotheses about confidence.

**Pierce’s Model of Confidence**

Pierce’s (1877) hypothesis was that confidence reflected Fechner’s perceived interval of uncertainty and as a result confidence should be logarithmically related to the chance of correctly detecting a difference between stimuli. More formally, Pierce and Jastrow (1884) empirically demonstrated that the average confidence rating in a discrimination task was well described by the expression:

$$\text{conf} = \beta \cdot \ln \left[ \frac{P(R_A | S_A)}{P(R_B | S_A)} \right]$$

(9)
The parameter $\beta$ is a scaling parameter. While innovative and thought provoking for its time, the law is descriptive at best. Link (1992; 2003) and Heath (1984), however, reformulated Equation 9 into the process parameters of the DSD model. If we assume no bias on the part of the judge then substituting the DSD choice probabilities (see Equation A1) into Equation 9 yields

$$
\text{conf} = \ln \left[ \frac{P(R_A \mid S_A)}{P(R_B \mid S_A)} \right] / 2 = \delta \theta / \sigma^2.
$$

(10)

In words, Pierce’s hypothesis implies confidence is a multiplicative function of the quantity of the information needed to make a decision ($\theta$; or the distance traveled by the diffusion process) and the quality of the information ($\delta$; or the rate of evidence accumulation in the diffusion process) accumulated in DSD (for a more general derivation allowing for response bias see Heath, 1984). For the remainder of this paper, this function in combination with a DSD model describing choice and decision time is called Pierce’s model. Link (2003) and Heath (1984) showed that Pierce’s model gave a good account of mean confidence ratings.

In terms of passing the empirical hurdles, Pierce’s model passes several of the hurdles and in fact identifies two new hurdles that any model of confidence should account for. Of course Pierce’s model using the DSD framework clears hurdle 1: the speed/accuracy tradeoff. Pierce’s model also accounts for the positive relationship between discriminability and confidence (hurdle 2) because as countless studies have shown the drift rate systematically increases as stimulus discriminability increases (e.g., Ratcliff & Rouder, 1998; Ratcliff & Smith, 2004; Ratcliff, Van Zandt, & McKoon, 1999). According to Pierce’s model (Equation 10), this increase in drift rate implies that mean confidence increases.

Notice, though, that Pierce’s function is silent in terms of hurdle 3 where confidence for correct choices is greater than for incorrect choices. This is because Pierce’ function uses correct and incorrect choice proportions to predict the mean confidence. More broadly any hypothesis positing confidence to be a direct function of the diffusion model parameters ($\delta, \theta, z$) will have difficulty predicting a difference between corrects and incorrect trials, because these parameters are invariant across correct and incorrect trials.

Despite this setback, Pierce’s model does bring to light two additional hurdles that 2DSD and any model of confidence must surmount. Pierce’s model predicts that there is a negative relationship between decision time and the degree of confidence expressed in the choice. This is because as drift rate decreases the average decision time increases while according to Equation
10 confidence decreases. This empirical prediction has been confirmed many times where across trials under the same conditions the average decision time monotonically decreases as the confidence level increases (e.g., Baranski & Petrusic, 1998; Festinger, 1943; D. M. Johnson, 1939; Vickers & Packer, 1982). This negative relationship between decision time and confidence in optional stopping tasks serves as empirical hurdle 4 in Table 1.

This intuitive negative relationship between confidence and decision time has been the bedrock of several alternative accounts of confidence that postulate judges use their decision time to form their confidence estimate, where longer decisions times are rated as less confident (Audley, 1960; Ratcliff, 1978; Volkman, 1934). These time-based hypotheses, however, cannot account for the positive relationship between decision time and confidence that Pierce’s model also predicts. That is, as the choice threshold (θ) or the quantity of information collected increases confidence should also increase.

Empirically a positive relationship between decision times and confidence was first identified in interrogation choice paradigms where an external event interrupts judges at different sample sizes and asks them for a choice. Irwin, Smith, and Mayfield (1956) used an expanded judgment task to manipulate the amount of evidence collected before making a choice and confidence judgment. An expanded judgment task externalizes the sequential sampling process asking people to physically sample observations from a distribution and then make a choice. As judges were required to take more observations, hence greater choice thresholds and longer decision times, their confidence in their choices increased (for a replication of the effect see Vickers, Smith, Burt, & Brown, 1985).

At the same time, when sampling is internal if we compare confidence between different levels of time pressure during optional stopping tasks, then we find that confidence is on average greater when accuracy is a goal as opposed to speed (Ascher, 1974; Vickers & Packer, 1982). In some cases, though, experimenters have found equal confidence between accuracy and speed conditions (Baransi & Petrusic, 1998; Festinger, 1943; Garrett, 1922; D. M. Johnson, 1939). This set of contradictory findings is an important limitation to Pierce’s model, which we will return to shortly. Regardless across these different tasks, this positive relationship between decision time and confidence eliminates many models of confidence and serves as hurdle 5 for any model of confidence.

In summary, although Pierce’s model appears to have a number of positive traits, it also
has some limitations as a dynamic account of confidence. The limitations by and large can be attributed to the fact that confidence in Pierce’s model is a direct function of the quantity ($\theta$) and quality ($\delta$) of the evidence used to make a choice. This assumption seems implausible because it implies a judge would have direct cognitive access to such information. If judges knew the drift rate they shouldn’t be uncertain in making a choice to begin with. But, even if this plausibility criticism is rectified in some manner, there is another more serious problem: Pierce’s model cannot clear hurdle 3 where judges are more confident in correct trials than incorrect trials. This limitation extends to any other model that assumes confidence is some direct function of the quantity ($\theta$) and quality ($\delta$) of the evidence. An alternative hypothesis is that at the time a judge enters a confidence rating, judges do not have direct access to the quantity ($\theta$) and/or quality ($\delta$) of evidence, but instead have indirect access to $\theta$ and $\delta$ via some form of the actual evidence they accumulated. 2DSD makes this assumption.

**The 2 Stage Dynamic Signal Detection Model (2DSD)**

Typically DSD assumes that when the accumulated evidence reaches the evidence states of $\theta_A$ or $\theta_B$ the process ends and the state of evidence remains in that state thereafter. In other words, judges stop accumulating evidence. 2DSD relaxes this assumption and instead supposes that a judge does not simply shut down the evidence accumulation process after making a choice, but continues to think about the two options and accumulates evidence to make a confidence rating (see Figure 1). Thus, the confidence rating is a function of the evidence collected at the time of the choice plus the evidence collected after making a choice.

There are several observations that support the assumptions of 2DSD. Neurophysiological studies using single cell recording techniques with monkeys suggest that choice certainty or confidence is based on the firing rate of the same neurons that also determines choice (Kiani & Shadlen, 2009). That is, confidence is a function of the state of evidence accumulation, $L(t)$. There is also support for the second stage of evidence accumulation. Anecdotally, we have probably all had the feeling of making a choice and then almost instantaneously new information comes to mind that changes our confidence in that choice. Methodologically some psychologists even adjust their methods of collecting confidence ratings to account for this post-decision processing. That is, after making a choice, instead of asking judges to enter the confidence that they are correct (.50, …, 1.00) (a two-choice half range method) they ask judges to enter their confidence that a pre-specified alternative is correct (.00,
Modeling choice, decision time, and confidence

..., 1.00) (a two-choice full range method) (Lichtenstein et al., 1982). The reasoning is simple: the full range helps reduce issues participants might have where they make a choice and then suddenly realize the choice was incorrect. But, more importantly, the methodological adjustment highlights our hypothesis that judges do not simply stop collecting evidence once they make a choice, but continue collecting evidence.

Behavioral data also support this notion of post decisional evidence accumulation. For instance, we know even before a choice is made that the decision system is fairly robust and continues accumulating evidence at the same rate even after stimuli are masked from view (Ratcliff & Rouder, 2000). Several results also imply that judges continue accumulating evidence even after making a choice. For instance, judges appear to change their mind even after they have made a choice (Resulaj, Kiani, Wolpert, & Shadlen, 2009). Furthermore, if judges are given the opportunity to enter a second judgment not only does their time between their two responses (inter-judgment time) exceed motor time (Baranski & Petrusic, 1998; Petrusic & Betrusic, 2003), but judges will sometimes express a different belief in their second response than they did at their first response (Van Zandt & Maldonado-Molina, 2004).

One way to model this post-decision evidence accumulation is with the interrogation model of 2DSD. In this version, after reaching the choice threshold and making a choice, the evidence accumulation process continues for a fixed period of time $\tau$ or inter-judgment time. In most situations, the parameter $\tau$ is empirically observable. Baranski and Petrusic (1998) examined the properties of inter-judgment time in a number of perceptual experiments involving choice followed by confidence ratings and found (a) if accuracy is stressed, then the inter-judgment time $\tau$ is between 500 to 650 ms and can be constant across confidence ratings (especially after a number practice sessions); (b) if speed is stressed, then $\tau$ was higher (~700 to 900 ms) and seemed to vary across confidence ratings. This last property (inter-judgment times varying across confidence ratings) suggests that the inter-judgment time is determined by a dynamic confidence rating judgment process. But, for the time being, we will assume that $\tau$ is an exogenous parameter in the model.

At the time of the confidence judgment, denoted $t_c$, the accumulated evidence reflects the evidence collected up to the decision time $t_d$, plus the newly collected evidence during the period of time $\tau = n\Delta t$.

$$ L(t_c) = L(t_d) + \sum_{i=1}^{n} x(t_d + i \cdot \Delta t) $$  \hspace{1cm} (11)
Analogous to signal detection theory (e.g., Macmillan & Creelman, 2005), judges map possible ratings onto the state of the accumulated evidence \(L(t_c)\). In our tasks there are six levels of confidence (\(conf = .50, .60, \ldots, \text{and } 1.00\)) conditioned on the choice \(R_A\) or \(R_B\), \(conf_j|R_i\) where \(j = 0, 1, \ldots, 5\). So each judge needs five response criteria for each option, \(c_{k,R_a}\) where \(k = 1, 2, \ldots, 5\) to select among the responses. The response criteria, just like the choice thresholds, are set at specific values of evidence. The locations of the criteria depend on the biases of judges. They may also be sensitive to the same experimental manipulations that change the location of the starting point, \(z\). For the purpose of this paper, we will assume they are fixed across experimental conditions and are symmetrical for \(R_A\) or \(R_B\) response (e.g., \(c_{k,R_a} = -c_{k,R_b}\)). Future research will certainly be needed to identify if and how these confidence criteria move in response to different conditions. With these assumptions, if judges choose the \(R_A\) option and the cumulated evidence is less then \(L(t_c) < c_{1,R_a}\) then judges select the confidence rating \(.50\), if it rests between the first and second criteria, \(c_{1,R_a} < L(t_c) < c_{2,R_a}\), then they choose \(.60\), and so on.

The distributions over the confidence ratings are functions of the possible evidence accumulations at time point \(t_c\). The distribution of possible evidence states at time point \(t_c\) in turn reflect the fact that we know what state the evidence was in at the time of choice, either \(\theta_A\) or \(\theta_B\). So our uncertainty about the evidence at \(t_c\) is only a function of the evidence accumulated during the confidence period of time \(\tau\). Thus, based on Equation 3 and assuming evidence is accumulated continuously over time (\(\Delta t \rightarrow 0\)), when stimulus \(S_A\) is present the distribution of evidence at time \(t_c\) is normally distributed with a mean of

\[
E[L(t_c)|S_A] = \begin{cases} 
\tau \delta + \theta_A, & \text{if } R_A \text{ was chosen} \\
\tau \delta - \theta_B, & \text{if } R_B \text{ was chosen}
\end{cases}
\]

The means for stimulus \(S_B\) trials can be found by replacing the \(\delta\)'s with \(-\delta\). The variance, following Equation 4, in all cases is

\[
\text{var}[L(t_c)] = \sigma^2 \tau
\]

The distribution over the different confidence ratings \(conf_j\) for hit trials (respond \(R_A\) when stimulus \(S_A\) is shown) is then

\[
\Pr\left( conf_j \mid R_A, S_A \right) = P\left( c_{j,R_a} < L(t_c) < c_{j+1,R_a} \mid \delta, \sigma^2, \theta_A, \tau \right)
\]
where $c_{0,R_A}$ is equal to $-\infty$ and $c_{8,R_A}$ is equal to $\infty$. Similar expressions can be formulated for the other choices. The precise values of $\Pr\left(\text{conf} \mid R_A, S_A\right)$ can be found using the standard normal cumulative distribution function. Table 2 lists the parameters of the 2DSD model. The total number of parameters depends in part on the number of confidence ratings.

**How does the model stack up against the empirical hurdles?** To begin notice that the means of the distributions of evidence at $t_C, L(t_C)$, are directly related to the drift rate and choice thresholds (Equation 12). Thus, 2DSD makes similar predictions as Pierce’s model, though Pierce’s model posits a multiplicative relationship as opposed to an additive one (Equation 10). The model still accounts for the speed-accuracy tradeoff (hurdle 1) because we use the standard diffusion model to make the choices. To explain why confidence is positively related to stimulus discriminability (hurdle 2), 2DSD relies on the fact that as stimulus discriminability increases so does the drift rate ($\delta$) and consequently confidence increases. The model can also correctly predict higher levels of confidence for accurate choices compared to incorrect ones (hurdle 3). To see why, notice that the mean of the evidence at the time confidence is selected is $\theta_A + \tau \delta$ for hits (response $R_A$ is correctly chosen when stimulus $S_A$ was shown) and $\theta_A - \tau \delta$ for false alarms (response $R_A$ is correctly chosen when stimulus $S_B$ was shown) (see Equation 12). In other words, the average confidence rating under most conditions will be greater for correct responses.

Similarly, decreases in the drift rate also produce longer stage 1 decision times and lower levels of confidence because confidence increases with drift rate. Thus, the model predicts a negative relationship between confidence and decision time (hurdle 4). The 2DSD model also predicts a positive relationship between confidence and decision times in both optional stopping and interrogation paradigms (hurdle 5). In optional stopping tasks, again judges set a larger threshold $\theta$ during accuracy conditions than in speed conditions. As a result this will move the means of the confidence distributions out producing higher average confidence ratings in accuracy conditions. During interrogation paradigms average confidence increases as judges are forced to accumulate more evidence (or take more time) on any given trial. Within the 2DSD model this implies that the expected state of evidence will be larger because it is a linear function of time (Equation 3) and thus confidence will be greater when judges are forced to take more time to make a choice.
Finally, 2DSD also accounts for a number of other phenomena. One example of this is an initially puzzling result where comparisons of confidence between speed and accuracy conditions showed that there was no difference on average confidence ratings between the two conditions (Festinger, 1943; Garrett, 1922; D. M. Johnson, 1939). This result speaks to some degree against hurdle 5. Vickers (1979) observed that when obtaining confidence ratings, participants are typically encouraged to use the complete range of the confidence scale. This combined with the fact that in previous studies speed and accuracy was manipulated between sessions prompted Vickers (1979) to hypothesize that participants spread their confidence ratings out across the scale within each session. As a result they used the confidence scale differently between sessions and this in turn would lead to equal confidence across accuracy and speed conditions. In support of this prediction Vickers and Packer (1982) found that when the “complete scale” instructions were used in tandem with manipulations of speed and accuracy within sessions, judges were less confident during speed conditions (though see Baranski & Petrusic, 1998). 2DSD naturally accounts for this result because it makes explicit – via confidence criteria – the process of mapping a confidence rating to the state of evidence at the time of the confidence rating.

An additional advantage of making the confidence mapping process explicit in 2DSD is that it does not restrict the model to a specific scale of confidence ratings. 2DSD can be applied to a wide range of scales long used in psychology to report levels of confidence in a choice such as numerical Likert type scales (“1”, “2”, …) to verbal probability scales (“guess”, …, “certain”) to numerical probability scales (“.50”, “.60”, …, “1.00”) (for a review of different confidence or subjective probability response modes see Budescu & Wallsten, 1995). This is a strong advantage of the model and we capitalize on this property later to connect the model to the decision sciences where questions of the accuracy of subjective probability estimates are tantamount.

Summary

We have presented a dynamic signal detection model where after making a choice, judges continue to accumulate evidence in support of the two alternatives. They then use the complete set of evidence to estimate their confidence in their choice. We have shown that this basic model accounts for a wide range of historical findings (hurdles 1 through 5 in Table 1). Accounting for these datasets is an important set of hurdles to clear because they have been used to rule out
Modeling choice, decision time, and confidence

possible alternative theories of confidence rooted within random walk/diffusion theory as well as many other theories (Vickers, 1979; Vickers 2001). 2DSD also makes a number of new predictions. For example, as Figure 1 and Equations 12, 13, and 14 imply, 2DSD does generally predict that – all else being equal – increases in \( \tau \) should increase both the mean difference between confidence ratings in correct responses and incorrect choices (slope) and the pooled variance of the distribution of confidence ratings across correct and incorrect choices (scatter).

These are strong predictions and at first glance intuition might suggest incorrect predictions. To test these predictions we used data collected in a new study where participants completed two often studied but different two-alternative forced choice situations: perceptual and general knowledge.

**Overview of empirical evaluation of 2DSD**

During the study, six participants completed a perceptual and a general knowledge task. In the perceptual task participants were shown 1 of 6 possible pairs of horizontal lines and asked to (a) identify which line was longer/shorter and then (b) rate their confidence in their response on a subjective probability scale (.50, .60, …, 1.00). This task has been studied in a number of studies on confidence and calibration (Baranski & Petrusic, 1998; Henmon, 1911; Juslin & Olsson, 1997; Vickers & Packer, 1982). During the conceptual task, participants were shown a pair of U.S. cities randomly drawn from the 100 most populated U.S. cities in 2006 and asked to identify the city with the larger/smaller population and then rate their response on a subjective probability scale. This is a common task studied repeatedly in studies on the accuracy of subjective probabilities (Gigerenzer, Hoffrage, & Kleinbölting, 1991; Juslin, 1994; McClelland & Bolger, 1994). Thus, it is of interest to replicate the studies with the goal of identifying the degree to which the same process can account for judgment and decision processes across these two tasks especially in light of claims that the two tasks draw on different processes (Dawes, 1980; Juslin, Winman, & Persson, 1995; Keren, 1988).

In both tasks, we manipulated the time pressure participants faced during the choice. When assessing confidence, however, participants were told to balance accuracy and speed. This allowed us to examine the effect of one type of time pressure on subjective probability estimates. Notably, 2DSD exposes a larger range of conditions where time pressure might influence the cognitive system. For example, a judge might face time pressure both when making a choice and when making a confidence rating or they might be in a situation where accurate choices and
confidence ratings are of utmost concern and any time pressure has been lifted altogether. In all cases, the model makes testable predictions. We chose, however, to focus on a situation when judges face time pressure when making a choice, but do not face severe time pressure when entering their confidence. Anecdotally at least, there are also a number real world analogs to our situation where judges typically have to make a quick choice and then retrospectively assess their confidence with less time pressure. For example, athletes must often make a split second choice on the playing field, military personnel have to make rapid decisions in a battle, or an investor must decide quickly to buy stock based on a tip. Later these agents face less time pressure when they are asked to assess their confidence that they made the correct choice. Nevertheless, future studies should certainly be designed to examine a broader range of time pressure across the different response procedures.

**Method**

**Participants**

The 6 participants were Michigan State University students. Five were psychology graduate students and one was an undergraduate student. Two were men and four were women. The participants were paid $8 plus a performance-based reward for their participant in each of the approximately 20 sessions. All the participants were right handed and had normal or corrected-to-normal vision. Participants earned between $10 and $14 for each session.

**Apparatus**

All stimuli were presented using software programmed in E-Prime 2.0 Professional. This allowed for controlled presentation of graphics, instructions, event sequencing, timing, and recording of responses. Participants recorded their responses using a standard Dell QWERTY keyboard with most of the keys removed save two rows. The first row contained the ‘V’, ‘B’, and ‘N’ keys and were relabeled as ‘←’, ‘H’, and ‘→’keys, respectively. Participants used the ‘H’ key to indicate their readiness for the two alternatives to be shown and then entered their choice with the respective arrow keys. After making a choice, participants immediately entered a confidence rating using the second row of keys. This row only contained the ‘D’ through ‘K’ keys and were relabeled with confidence ratings ‘50’, … ‘100’ to correspond with the confidence ratings in percentage terms. Participants were instructed to use their dominant hand to enter all responses. Periodic inspections during each session confirmed that participants adhered strictly
to this instruction. Participants sat in individual sound attenuated booths approximately 60 cm away from the screen.

**Response entry task**

Before the two-alternative forced choice experimental task, participants completed a response entry task where they entered a sequence of responses (e.g., ‘H’, ‘←’, ‘60’). The task helped participants practice the choice and confidence key locations. We also used the task to examine the degree to which there was any relationship between motor times and button locations. During the task, a screen instructed participants to press a sequence of responses (e.g., ‘←’ then ‘60’). When they were ready they were instructed to press the ‘H’ key and then as quickly as possible the appropriate choice key ‘←’ or ‘→’ then the confidence key. Participants entered each response sequence (choice and confidence) twice for a total of 24 trials. Accuracy was not enforced. However, due to a programming error during the line length experimental sessions, participants 1 through 4 were forced to enter the correct button. These participants completed an extra set of 30 trials per response sequence during an extra session. There was no systematic difference between the motor times associated with pressing the different choice keys. Across participants the average time to press the choice key was 0.221 s ($Std_{bwn} = 0.032$; $SE = 0.003$). There was also no systematic difference among the motor times for pressing the confidence buttons. The average time for pressing the confidence key after pressing a choice key was 0.311 s ($Std_{bwn} = 0.066$; $SE = 0.003$).

**Line length discrimination task**

**Stimuli.** The stimulus display was modeled after Baranski and Petrusic’s (1998) horizontal line discrimination tasks. The basic display had a black background. A white 20 mm vertical line in the center served as a central maker placed in the center of the screen. Two orange horizontal line segments extended to the left and right of the center line with 1 mm space between the central line and the start of each line. All lines including the central line were approximately 0.35 mm wide. In total there were six different pairs of lines. Each pair consisted of a 32 mm standard and the comparison was a 32.27, 32.59, 33.23, 33.87, 34.51, or 35.15 mm line.

**Design and procedure.** Participants completed 10 consecutive experimental sessions for the line length discrimination task. During each session participants completed three tasks. The first task was the previously described response entry task. The second task was a practice set of
8 trials (a block of 4 accuracy trials and a block of 4 speed trials counterbalanced across sessions). The order of practice blocks was counterbalanced across participants and between sessions. The final task was the experimental task. During the experimental task, participants completed eight blocks of trials. During each block (speed or accuracy) participants completed 72 trials (6 line pairs x 2 presentation orders x 2 longer/shorter instructions x 3 replications). The longer/shorter instructions meant that for half the trials participants were instructed to identify the longer line and for the other half they were instructed to identify the shorter line. This set of instructions replicates conditions used by Baranski and Petrusic (1998), Henmon (1911), and others in the foundational studies of confidence. Half the participants began session 1 with an accuracy block and half began with a speed block. Thereafter, participants alternated between beginning with a speed or accuracy block from session to session. In total participants completed 2,880 accuracy trials and 2,880 speed trials.

Participants were told before each block of trials their goal (speed or accuracy) for the upcoming trials. Furthermore, throughout the choice stage of each trial they were reminded of their goal with the words speed or accuracy at the top of the screen. During the accuracy trials, participants were instructed to enter their choice as accurately as possible. They received feedback after entering their confidence rating when they made an incorrect choice. During the speed trials participants were instructed to try and enter their choice quickly, faster than 750 ms. Participants were still allowed to enter a choice after 750 ms, but were given feedback later after entering their confidence that they were too slow in entering their choice. No accuracy feedback was given during the speed conditions. Participants were instructed to balance between entering an accurate, but quick confidence rating.

An individual trial worked as follows. Participants were first given a preparation slide which showed (a) the instruction (shorter or longer) for the next trial in the center; (b) a reminder at the top of the screen of the goal during the current block of trials (speed or accuracy); and (c) the number of trials completed (out of 72) and the number of blocks completed (out of 8). When they were ready, participants pressed the ‘H’ key, which (1) removed trial and block information; (2) moved the instruction to the top of the screen; and (3) put a fixation cross in the center. Participants were told to fixate on the cross and press the ‘H’ key when ready. This press removed the fixation cross and put the pair of lines in the center with the corresponding choice keys at the bottom (‘←’ or ‘→’). Once a choice was entered a confidence scale was placed
below the corresponding keys and participants were instructed to enter their confidence that they chose the correct line (50, 60, … 100%). After entering a confidence rating, feedback was given if they made an incorrect choice in the accuracy block or if they were too slow in the speed block. Otherwise no feedback was given and participants began the next trial.

At the beginning of the experiment, participants were told to select a confidence rating so that over the long run the proportion of correct choices for all trials assigned a given confidence rating should match the confidence rating given. Participants were reminded of this instruction before each session. This instruction is common in studies on the calibration of subjective probabilities (cf., Lichtenstein et al., 1982). As further motivation, participants earned points based on the accuracy of their choice and confidence rating according to the quadratic scoring rule,

$$\text{points} = 100 \left[ 1 - \left( \text{correct}_i - \text{conf}_i \right)^2 \right]$$

(Stäel von Holstein, 1970). Where \( \text{correct}_i \) is equal to 1 if the choice on trial \( i \) was correct otherwise 0 and \( \text{conf}_i \) was the confidence rating entered in terms of probability of correct (.5, .6, …, 1.0). This scoring rule is a variant of the Brier score (Brier, 1950) and as such is a strictly proper scoring rule insuring that participants will only maximize their earnings if they maximize their accuracy in both their choice and their confidence rating. Participants were informed of the properties of this scoring rule prior to each session and shown a table demonstrating why it was in their best interest to accurately report their choice and confidence rating. To enforce time pressure during the speed conditions, the points earned were cut in half if a choice exceeded the deadline of 750 ms and then every 500 ms after that the points were cut by half again. For every 10,000 points they earned one additional $1.

**City population discrimination task**

*Stimuli.* The city pairs were constructed from the 100 most populated U.S. cities from the 2006 U. S. Census estimates. There are 4,950 pairs. From this population, ten experimental lists of 400 city pairs were randomly constructed (without replacement). The remaining pairs were used for practice. During the choice trials, the city pairs were shown in the center of a black screen centered around the word “or” in red and written in yellow. Immediately below each city was the state abbreviation (e.g., MI).

*Design and procedure.* The city population task worked much the same way as the line discrimination task except the practice trials preceded the response entry task. The practice was
structured the same way as the line discrimination task with one block of four accuracy trials and one block of four speed trials. During the experimental trials participants again alternated between speed and accuracy blocks of trials with each block consisting of 50 trials. Half the trials had the more populated city on the left and half on the right. Half of the trials instructed participants to identify the more populated city and the other half the less populated city. Half the participants began session 1 with an accuracy block and half began with a speed block. Thereafter, participants alternated between beginning with a speed or accuracy block from session to session. In total participants completed 2,000 speed trials and 2,000 accuracy trials. Due to a computer error, Participant 5 completed 1,650 speed trials and 1,619 accuracy trials.

Instructions and trial procedures were identical across tasks. The only difference was that the deadline for the speed condition was 1.5 seconds. Pilot testing revealed that this was a sufficient deadline that allowed participants to read the cities, but insured they still felt sufficient time pressure to make a choice. Participants 1 to 4 completed the line length sessions first and then the city population sessions. Participants 5 and 6 did the opposite order. The study was not designed to examine order effects, but there did not appear to be substantial order effects.

**Results**

The results section is divided into four sections. The first section summarizes the behavioral results from the two tasks and examines several qualitative predictions that 2DSD makes regarding the effect of changes in inter-judgment time $\tau$ on confidence. These predictions are also of interest because race models using Vicker’s (1979) balance of evidence hypothesis of confidence predict the opposite pattern of results. The second section examines the fit of the 2DSD interrogation model to the data. In this section, we also investigated the degree to which trial variability in the process parameters – a construct of interest to both cognitive (Ratcliff & Rouder, 1998) and decision scientists (Erev, Wallsten, & Budescu, 1994) alike – adds to the fit of the 2DSD model. In the third section, we provide an explanation for the trade-offs made between the entire time course of the judgment process (decision time + inter-judgment time) and the accuracy of both the observed choice and confidence rating. This tradeoff is not only important for decision scientists who have long focused on the accuracy of confidence judgments, but the tradeoff can help explain the observed behavior of participants in our study. Finally, we present an evaluation of a more extensive version of 2DSD that offers a more precise process account of the distributions of inter-judgment times.
As a first step in the data analyses, in both the line and city tasks, we removed trials that were likely the result of different processes thus producing contaminant response times (Ratcliff & Tuerlinckx, 2002). To minimize fast outliers, we excluded trials where decision times were less than 0.3 seconds and the observed inter-judgment times were less than 0.15 seconds. To minimize slow outliers, we excluded trials where either the decision time or observed inter-judgment time was greater than 4 standard deviations from the mean. These cutoffs eliminated on average 2.5% (min = 1.1%; max = 4.9%) of the data in the line task and 2.0% (min = 1.0%; max = 5.3%) of the data in the city task.

**Behavioral tests of 2DSD**

**Between condition results.** Table 3 lists the proportion correct, the average decision time, the average confidence rating, and the average inter-judgment time for each participant in the line length and city population task. The values in the parentheses are standard deviations. The far right column lists the average statistic across participants. Throughout the paper when statistics are listed averaged across participants they were calculated using methods from meta-analysis where each participant’s data was treated as a separate experiment and the average statistic is calculated by weighting each participant’s respective statistic by the inverse of the variance of the statistic (Shadish & Haddock, 1994). These estimates were calculated assuming a random effects model. Calculating the average statistic in this way provides an estimate of the standard error around the average statistic.

The descriptive statistics reveal that, by and large, the time pressure manipulation worked. Focusing first on the decision stage, in the line length task for all six participants both the proportions correct and decision times were significantly smaller in the speed condition (hurdle 1 in Table 1). A similar pattern emerges for the city population task. In terms of the confidence judgments, all six participants were significantly less confident in the speed condition for both the line and city tasks. In other words, consistent with hurdle 5, there is a positive relationship between confidence and decision time between conditions. The primary explanation for this decrease in the average confidence ratings again rests, everything else remaining equal, with the decrease in the magnitude of the choice threshold \( \theta \) (Equation 12).

All else, however, did not remain equal. In fact, as Table 3 also shows judges on average increased their inter-judgment time during the speed conditions. That is, judges appear to compensate for their time pressure when making a choice by taking a little longer to rate their
Modeling choice, decision time, and confidence. This is true for both the line and city tasks. Similar results are reported in Baranski and Petrusic (1998). According to the 2DSD model, this increase in inter-judgment time (and thus additional evidence accumulation) moderates the degree to which the change in thresholds can account for the decrease in average confidence. We will return to the implications of this interaction between changes in the choice threshold and inter-judgment time shortly.

More broadly, though, we interpret this result of increased inter-judgment time during the speed conditions as offering preliminary support for our hypothesis that judges continue to engage in post decisional stimulus processing to enter their confidence rating. If no post decisional processing occurred and instead choice and confidence are simultaneously available as most models assume, then one would expect no difference in inter-judgment time between time pressure conditions. In the third section, we show that this strategy of increasing inter-judgment time may be optimal in terms of producing the most accurate choice and confidence rating in the least amount of time. Before we examine any of these implications it is also useful to consider the within-condition relationships between the various measures of cognitive performance in the two tasks.

Within condition results. To evaluate the within condition relationships between the cognitive performance measures, we used the Goodman and Kruskal’s $\Gamma$ ordinal measure of association (Goodman & Kruskal, 1953, 1954). Goodman and Kruskal’s $\Gamma$ only assumes an ordinal scale and makes no distribution assumptions (Goodman & Kruskal, 1953). In addition, unlike Pearson’s $r$, $\Gamma$ can attain its maximum value regardless of the presence of numerical ties in either of the two correlated variables (Gonzalez & Nelson, 1996). This is especially relevant when ties in a particular variable (confidence ratings) are not necessarily theoretically meaningful (Nelson, 1984, 1987).

Table 4 lists the Goodman and Kruskal’s $\Gamma$ between the measures of cognitive performance in each of the tasks averaged across participants. The values below the diagonal are the $\Gamma$ coefficients for the accuracy condition and above the diagonal are the values for the speed condition. The values in parentheses are an estimate of the between participant standard deviation of the $\Gamma$ coefficient. The associations listed in Table 4 are in line with the empirical hurdles laid out in Table 1. Focusing on confidence, there is a monotonic relationship between confidence and an objective measure of difficulty in both tasks (hurdle 2). Difficulty in the line length task is the difference between line lengths whereas difficulty in the city population task is
indexed by the difference between the ordinal ranking of the cities. This latter measure of
difficulty in the city population task is based on the idea that the quality of information stored in
memory is often related (either directly or via mediators) to relevant environmental criteria (like
city populations) (Goldstein & Gigerenzer, 2002). Table 4 also shows that confidence is
monotonically related to accuracy (hurdle 3) and inversely related to decision time (hurdle 4).  

Notice also the pattern of Γ’s is largely consistent across the two tasks, though the
magnitude of the associations is smaller in the city population task. Presumably this decrease in
magnitude is due to a larger amount of variability both from the stimuli and the participant.
Nevertheless, the pattern of associations is at least consistent with the hypothesis that although
the information comes from a different source, a similar decision process is used in each task.
The associations in Table 4 also reveal places where the 2DSD interrogation model is silent. In
particular, in all conditions and all tasks inter-judgment time and the confidence rating were
negatively related (Γ = -.16 to -.52), we will return to this result in the final results section.

The associations in Table 4 are also revealing in terms of the accuracy of confidence.
Namely in both tasks and in both conditions one of the largest correlations was between accuracy
and the confidence rating (Γ = .43 to .75). Thus, participants exhibited good resolution in their
confidence ratings (Nelson, 1984). Also of interest is the change in resolution between time
pressure conditions. In fact, the average participant had better resolution during the speed
conditions in both the line length (Γ = .75 vs. .67, p < .01) and city population (Γ = .54 vs. .43, p
< .01). The 2DSD interrogation model attributes this increased resolution of confidence during
time pressure to the increase in inter-judgment time. Next we analyze this prediction in better
detail.

**Changes in the distribution of confidence ratings under time pressure.** Everything else
being equal, 2DSD predicts that as inter-judgment time τ increases the mean and the variance of
the distribution of evidence used to estimate confidence $L(t_c)$ should increase (see Equation 12
Equation 13). These changes in the average state of evidence imply that the slope score (Yates,
1990) should increase as inter-judgment time τ increases. Slope is calculated according to the
expression

$$slope = \overline{\text{conf}}_{\text{correct}} - \overline{\text{conf}}_{\text{incorrect}}.$$  (16)
The slope score is an unstandardized measure of resolution. It is so named because if we used linear regression to predict the confidence rating and the dichotomous variable of correct/incorrect is entered as a predictor, then the slope of the regression would be the slope score (Yates, 1990). Table 5 lists the slope score for each participant in the speed and accuracy conditions of both the line length and city population discrimination task. The slope statistics show that for most participants (5 out of 6) as well as the average participant, slope was significantly larger during the speed condition as compared to the accuracy condition. In other words, the confidence participants expressed had better unstandardized discrimination between correct and incorrect choices during the speed conditions as opposed to the accuracy conditions.

The increase in slope is consistent with the predictions of the 2DSD interrogation model when inter-judgment times $\tau$ increased. Note first the increase in slope occurred despite the presumably lower choice thresholds in the speed conditions. According to the model, lower choice thresholds ($\theta$) in the speed conditions lead to a decrease in the average confidence for both corrects and incorrects. Recall, however, increases in inter-judgment time $\tau$ lead to an increase in the confidence for corrects and a decrease in the confidence for incorrects. Thus, the combined effect of lower choice thresholds ($\downarrow \theta$) and greater inter-judgment times ($\uparrow \tau$) in the speed condition produce (a) a small change in the confidence for corrects between speed and accuracy and (b) a substantial decrease in the average confidence for incorrects. Indeed empirically this was the case. In the line length task the average confidence in corrects went from $.93 (SE = .04; Std_{bwn} = .05)$ in the accuracy condition to $.90 (SE = .04; Std_{bwn} = .07)$ in the speed condition, $p < .05$. In comparison, the average confidence in incorrects went from $.81 (SE = .04; Std_{bwn} = .11)$ to $.71 (SE = .05; Std_{bwn} = .09)$, $p < .01$. A similar pattern occurred in the city population task. The average confidence in corrects went from $.82 (SE = .04; Std_{bwn} = .10)$ in the accuracy condition to $.80 (SE = .04; Std_{bwn} = .10)$ in the speed condition, $p < .01$. In comparison, the average confidence in incorrects went from $.74 (SE = .05; Std_{bwn} = .08)$ to $.68 (SE = .05; Std_{bwn} = .10)$, $p < .01$. Thus, without fitting the 2DSD model to the data, the complex changes in confidence between speed and accuracy conditions are at least consistent with the model.\(^7\)

Another relevant statistic is the scatter score (Yates, 1990) or the pooled variance of confidence across the correct and incorrect choices,
According to 2DSD scatter should increase with longer inter-judgment times (τ) because the variance of the distribution of evidence at the time of confidence increases with inter-judgment time (Equation 13). Table 5 lists the scatter score for each participant in the speed and accuracy conditions of both the line length and city population discrimination task. The scatter statistics show that for most participants (4 out of 6 in the line length task and 5 out of 6 in the city population task) as well as the average participant, variance was significantly larger during the speed condition as compared to the accuracy condition. Note unlike the slope score, changes in the choice threshold θ have little to no effect on the scatter of participants because θ only influences the mean confidence level not the deviations from the mean.

In sum, we found that choices under time pressure were compensated by longer inter-judgment times. Under these conditions, the 2DSD model makes the counterintuitive predictions that time pressure increases both the slope and the scatter of the confidence ratings. We will return to these results later when we examine whether standardized resolution increases in the speed condition. A limitation of this prediction is that it does predict that if τ is large enough the model predicts that confidence must fall into one of the extreme categories (.50 and 1.00). This is probably an incorrect prediction and can be partly addressed by adding trial variability to the process parameters as we do in a later section or by assuming some decay in the evidence accumulation process (Bogacz et al., 2006; Busemeyer & Townsend, 1993). Nevertheless, the prediction that slope and scatter increase under time pressure is particularly important because a competing class of sequential sampling models for choice and confidence cannot predict this pattern of results.

Race models. Race models using Vicker’s (1979) balance of evidence hypothesis actually predict a decrease in slope and scatter as time pressure increases. This class of models offers an alternative sequential sampling process to choice. The basic idea is that when judges are presented with a choice between two alternatives evidence begins to accrue on counters, one for each response alternative. The first counter to reach a threshold determines the choice. Thus, choice in these models is based on absolute count of evidence as opposed to a relative amount of evidence as DSD models (Ratcliff & Smith, 2004).
Models using this type of sequential sampling process include the linear ballistic accumulator model (Brown & Heathcote, 2008), the Poisson race model (Pike, 1973; Townsend & Ashby, 1983), the accumulator model (Vickers, 1979), and other biologically inspired models (Usher & McClelland, 2004; Wang, 2002). They appear to give a good account of choice data though some models like the linear ballistic accumulator appear to do better at accounting for decision time distributions and changes in the distributions than others like the accumulator and Poisson race model (Brown & Heathcote, 2008; Ratcliff & Smith, 2004).

To model confidence with race models, Vickers (1979; 2001) proposed the balance of evidence hypothesis where confidence is “the difference between the two totals (on the counters) at the moment a decision is reached or sampling terminated” (Vickers, 2001, p. 151). Vickers and colleagues (Vickers, 1979, 2001; Vickers & Smith, 1985; Vickers, Smith et al., 1985) have shown that race models paired with the balance of evidence hypothesis can account for empirical hurdles 1 to 5 listed in Table 1. Moreover, using the Poisson race model Van Zandt (2000b) and Merkle and Van Zandt (2006) have shown that the balance of evidence hypothesis can also account for effects of response bias on changing the slope of ROC curves (Van Zandt, 2000b) and overconfidence (Merkle & Van Zandt, 2006).

Race models using a one-stage balance of evidence hypothesis, however, cannot account for the changes in the confidence distributions as a result of changes in time pressure at choice. This is because time pressure at choice causes a decrease in the total amount of evidence that can be collected on both counters. This decrease is due to either lower choice thresholds or a higher initial starting point of the counters, or both (cf., Bogacz, Wagenmakers, Forstmann, & Nieuwenhuls, 2010). In terms of the balance of evidence hypothesis, the decrease means that there will be both fewer possible differences and the magnitude of the possible differences will lower. Thus, under time pressure race models using the balance of evidence hypothesis predict that the mean confidence and variance of the confidence will shrink and consequently slope and scatter should decrease. Obviously, this pattern of results was not observed (Table 5).

A second stage of processing could be adopted in race models to account for these results. For instance, Van Zandt and Maldonado-Molina (2004) have developed a two-stage Poisson race model. Though we point out that the Poisson race model has several other difficulties it must overcome. For instance, the model has difficulty in accounting for distributions of decision times (Ratcliff & Smith, 2004). There are other race models that might
be better candidates for the two-stage balance of evidence hypothesis including Brown and Heathcote’s (2008) linear ballistic accumulator and Usher and McClelland’s (2004) leaky accumulator model. We leave a full model comparison with these models for future investigation. Next we examine several new quantitative predictions 2DSD makes. To do so it is useful to fit the 2DSD model to the multivariate distribution of choice, decision time, and confidence.

**Fit of 2DSD Interrogation Model to Choices, Decision Times and Confidence Ratings**

*Model estimation.* We fit the 2DSD interrogation model to the distributions of choices, decision times, and confidence ratings. To do so we adapted Heathcote, Brown, and Mewhort’s (2002) quantile maximum probability (QMP) estimation method to simultaneously fit the multivariate distribution of choice by decision time by confidence ratings (see also Speckman & Rouder, 2004). The general idea of QMP is to summarize the distribution of decision times in terms of quantiles. For example, we used quantiles of .1, .3, .5, .7, and .9. These estimated quantiles divide the observed distribution of decision times into bins (e.g., 6 bins) each with a number of observations within each bin (e.g., .1*n in the first bin and .2*n in the second bin, etc.). If the decision times were the only variable to be fit then the cumulative distribution function of DSD (see Appendix A) could be used to fit the model to the observed number observations within each quantile bin. This can be done for the correct and incorrect distribution of decisions times to obtain parameter estimates of the model.

The QMP method can be easily adapted to include confidence ratings. In our case, instead of six bins of observations the data representation is a 6 (decision time categories) x 6 (confidence ratings) data matrix for corrects and incorrects for both tasks. We fit the 2DSD interrogation model to this data representation with a multinomial distribution function using probabilities generated by the 2DSD interrogation model. This was done for both the line length and city population discrimination task. The models were fit at the individual level.

In both the line length and city population tasks we broke the data down into two levels of time pressure crossed with different levels of difficulty. Recall, during the line length discrimination task, within each time pressure condition, participants saw each of the six different comparisons 480 times. Unfortunately, all participants were extremely accurate with the sixth and easiest comparison (35.15 mm vs. a 32 mm standard). They scored 94% correct in speed and 98% correct in accuracy in this condition resulting in very few incorrect trials. As a
result we collapsed the sixth and fifth levels of difficulty for both speed and accuracy forming five levels of difficulty in the line length task to model. Representing the line length discrimination data in terms of our adapted QMP method, each condition had 71 free data points producing a total of $71 \times 10$ conditions ($2$ time pressure levels x $5$ levels of difficulty) = 710 data points per participant. In the city population task, based on the relationship between the cognitive performance variables and the rank difference between city populations within each pair (Table 4), we formed six different levels of difficulty with approximately $300$ pairs in each condition. Thus, the city task with $12$ conditions (speed vs. accuracy X $6$ levels of difficulty) had $71 \times 12 = 852$ free data points per participant.

**Trial variability and slow and fast errors.** To better account for the data, we incorporated trial-by-trial variability into some of the process parameters of 2DSD. In particular, DSD models with no starting point bias ($z = 0$) and without trial variability predict that the distribution for correct and incorrect choices should be equal (Laming, 1968; Link & Heath, 1975; Ratcliff & Rouder, 1998; Ratcliff et al., 1999; Townsend & Ashby, 1983). Empirically this is not observed. Slow errors are often observed especially when accuracy is emphasized during more difficult conditions. Slow errors have become an important hurdle for any model of decision times to overcome (Estes & Wessel, 1966; Luce, 1986; Swensson, 1972; Townsend & Ashby, 1983). We have added the result of slow errors as the sixth empirical hurdle in Table 1 that any complete models of cognitive performance must explain. Sometimes during the easier conditions – especially when time pressure is high – the opposite pattern of decision times is present where mean decision times of incorrect choices are faster than the mean decision time for correct choices (Ratcliff & Rouder, 1998; Swenson & Edwards, 1971; Townsend & Ashby, 1983). We have also added this result as hurdle 7 to Table 1.

These two hurdles – slow errors for difficult conditions (hurdle 6) and fast errors for easy conditions (hurdle 7) – are problematic for many sequential sampling models to explain. Ratcliff and colleagues have shown that in order to simultaneously account for this pattern of results DSD require trial-by-trial variability in the process parameters (Ratcliff & Rouder, 1998; Ratcliff et al., 1999; Ratcliff & Smith, 2004). Trial-by-trial variability in processing stimuli is an often made assumption in stochastic models. This is true not only for DSD, but also Thurstonian scaling (Thurstone, 1927) and signal detection theory (Green & Swets, 1966; Wallsten & Gonzáles-Vallejo, 1994). In terms of DSD and 2DSD one source of trial variability, perhaps due
to lapses in attention or variability in memory processes (Pleskac, Dougherty, Rivadeneira, & Wallsten, 2009), is in the quality of the evidence accumulated or the drift rate $\delta$. In particular, we assume that the drift rate for stimuli in the same experimental conditions is normally distributed with a mean $\nu$ and a standard deviation $\eta$. Variability in the drift rate across trials causes the diffusion model to predict slower decision times for incorrect choices than for correct choices (hurdle 6) (Ratcliff & Rouder, 1998).

A second source of trial variability is in the starting point ($z$). This assumption captures the idea that perhaps due to sequential effects from trial to trial judges do not start their evidence accumulation at the exact same state of evidence. In DSD and 2DSD this variability is modeled with a uniform distribution mean $z$ and range $s_z$. Variability in the starting point across trials leads DSD to predict faster decision times for incorrect choice than for correct choices, especially when speed is emphasized (Ratcliff & Rouder, 1998; Ratcliff et al., 1999). While this pattern is not strongly present in our data, we have also included this assumption of trial variability in the start point when fitting the 2DSD interrogation model.

A complete description of the 2DSD interrogation model parameters is given in Table 2. For each decision task, we fit a highly constrained model to each participant’s data. Except where explicitly noted, all parameters were fixed across conditions. Furthermore, no systematic response bias was assumed so that the distribution of starting points was centered at $z = 0$ and the choice thresholds were set equal $\theta = \theta_A = \theta_B$ (recall the lower threshold is then $-\theta$). Across the different levels of difficulty only the mean drift rate $\nu$ was allowed to vary. Thus, for each participant there were five mean drift rates $\nu$ for the line length task and six $\nu$s in the city task. For both decision tasks only the choice thresholds $\theta$ were allowed to vary between the two time pressure conditions. As an estimate of $\tau$ we used the observed mean inter-judgment time for the particular condition, corrected for any non-judgment related processing time $t_{EJ}$. With six confidence ratings (.50, .60, .70, .80, .90, and 1.00) there are five confidence criteria. The confidence criteria were also fixed across experimental conditions and were symmetrical for $R_A$ or $R_B$ response (e.g., $c_{k,R_B} = -c_{k,R_A}$). Thus, in the line length task the 2DSD interrogation model had a total of 16 free parameters to account for 710 data points per participant. In the city population task there were a total of 17 free parameters to account for 852 data points per participant. The estimated parameters for each person and task are shown in Table 6.
To evaluate the fit of the 2DSD interrogation model we adapted the latency-probability function plots first used by Audley and Pike (1965) to include confidence as well. We call the plots latency-confidence-choice probability (LCC) functions. Figure 2 shows the plots for the average participant during the line length task (top row) and the corresponding plots for the city population task (bottom row). Overall each LCC function can be understood as a plot showing how the measures of cognitive performance change with stimulus discriminability. Within each figure, the lower half of the figure plots the mean decision times against the choice probability for correct choice (grey circles right of choice probability .5) and incorrect choices (white circles left of choice probability .5) for each level of difficulty. Because the sequential sampling models were fit to five levels of difficulty in the line-length task there are five correct (grey) and five incorrect (white) circles. In the city population task there are six correct (grey) and incorrect (white) dots. The grey dot furthest to the right in each panel corresponds to the choice proportion and corresponding decision time for the easiest condition. In other words, it has the highest choice probability and typically the fastest decision time among corrects. The grey dot closest to the .5 choice probability corresponds to the proportion correct in the most difficult condition and the slowest average decision time among corrects. The white dots are configured similarly for incorrect choices with the easiest condition on the far left. The upper portion of each panel plots the mean confidence against the choice probability in the same fashion. The solid lines marked with squares correspond to the predicted functions for the 2DSD interrogation model. Recall while the LCC functions plot the means, the models were fit using the quantiles of the decision times.

In terms of confidence, consistent with the slope scores, the observed average confidence rating for corrects is greater than for incorrects. Furthermore, this difference between the confidence ratings (slope) increases as the stimuli get easier (moving out from the confidence-choice probability functions). This is true for both the line length (top row) and city population (bottom row) discrimination tasks. The 2DSD model gives a good account of this trend. Furthermore, the LCC functions also show that the positive slope relating confidence and accuracy was greater for the speed conditions as compared to accuracy conditions. The 2DSD model reproduces this change in slope. As discussed earlier, 2DSD attributes the change in confidence with changes in difficulty to changes in the mean drift rate, while the changes between time pressure conditions are attributed to an interaction between the inter-judgment time
and choice thresholds. The LCC functions do reveal that the model tends to underestimate the mean confidence for incorrect choices. One way to correct for this underestimation is to assume separate confidence criteria for correct and incorrect choices – we assumed symmetry to keep the model as simple as possible.

With respect to mean decision time, Figure 2 shows that although the model captures the pattern of changes across conditions, there is a constant discrepancy between the predicted and observed mean times for the speeded conditions. Under the constraints used to fit the 2DSD model, the predicted time systematically overestimates the observed time under the speeded conditions by an approximately constant amount. This is very likely caused by our assumption that the residual motor time, $t_{ED}$, is equal for speed and accuracy conditions. This constant error could be fixed by allowing a smaller residual time constant for the speeded conditions.

The LCC functions show that slow errors occurred where decision times were slower for errors than correct responses for the corresponding level of difficulty. Consistent with hurdle 6, the slow errors are especially evident when accuracy was emphasized and during the more difficult conditions. The 2DSD interrogation model with trial variability in the drift rate helps 2DSD account for this pattern of slow errors. Without trial variability the model predicts identical decision time distributions for correct and incorrect choices (assuming no bias) and the decision times in the LCC would be a symmetrical inverted U shape.

Trial variability in 2DSD also helps account for other phenomena. In particular, trial variability in the drift rate helps account for the observed relationship between confidence and decision times even when stimulus difficulty is held constant. Past studies (Baranski & Petrusic, 1998; Henmon, 1911) indicate that even when the stimulus difficulty is held constant (i.e., respond to repeated trials of the same line pairs) there is a negative relationship between decision time and confidence so that the fastest decision times are generally associated with the highest confidence rating. Indeed participants in the line task tended to show this pattern of results. Table 7 lists the average Goodman and Kruskal $\Gamma$ rank order correlation between decision time and confidence for each participant in the line length task holding the stimulus difficulty constant. These correlations reveal that for nearly everyone, holding stimulus difficulty constant, there was a negative correlation between decision time and confidence and the strength of this relationship was strongest for correct responses during the accuracy conditions.
The 2DSD model (without trial variability) predicts that for a given stimulus, decision times and confidence are independent of each other (separately for correct and incorrect choices). The 2DSD model with trial variability attributes this relationship between confidence and decision time to trial-by-trial variability in the drift rate. That is, even when stimulus discriminability is held constant there is a negative relationship between decision time and confidence (hurdle 4 in Table 1) because from trial to trial there are fluctuations in the processing of the same stimuli that lead to changes in the quality of the evidence being accumulated. On some trials when the quality of the evidence extracted from a stimulus is high the drift rate will be high and lead to fast decision times and high levels of confidence. On other trials when the quality of the evidence is low the drift rate will be low and lead to slow decision times and lower levels of confidence. The model does, however, underestimate the relationship between observed decision time and confidence. For example, as listed in Table 7 the average observed Goodman and Kruskal $\Gamma$ between decision time and confidence across participants and difficulty levels was -.21 and -.39 for correct choices in speed and accuracy conditions, respectively. The average predicted correlation generated from the best fitting model parameters for each participant for these conditions was -.10 and -.15, respectively. Nevertheless, this result provides further support for the necessity of including trial variability in the process parameters to give a full account of the data.

**Accuracy of confidence ratings**

So far we have shown the 2DSD model can explain the differences in confidence for speed and accuracy conditions by the longer inter-judgment time that the judges used under the speed condition. But why did they do this? An answer to this question comes from understanding how the time course of the judgment process impacts the accuracy of confidence judgments. In this section we first describe some of the basic measures of confidence accuracy and examine the predictions of the 2DSD model for these measures. Then we examine the question of how to optimize both choice and confidence accuracy under time pressure conditions.

Confidence is useful not only in the lab to help chart cognitive processes, but also outside of the lab where it is often communicated as a subjective probability that an event has occurred or will occur (Adams & Adams, 1961; de Finetti, 1962; Savage, 1954). An important and well-studied aspect of subjective probabilities is their accuracy (see Arkes, 2001; Griffin & Brenner, 2004; Koehler, Brenner, & Griffin, 2002; McClelland & Bolger, 1994). The focus on accuracy is
well-deserved. Every day many important decisions are made using subjective probabilities to weigh the costs and benefits of the consequences of those decisions. Yet, very little is known about how time pressure affects the time allocation for making both choices and confidence ratings. In fact, 2DSD can be used to address this problem, but to do so it is much easier to rely on the simpler 2DSD interrogation model. As we have shown, the 2DSD interrogation model by treating inter-judgment time as an exogenous parameter captures all of the critical phenomena regarding the relationship between choice, decision time, and confidence.

**Substantive goodness.** There are, in fact, two different dimensions by which the accuracy of subjective probability estimate can be evaluated: substantively and normatively (Winkler & Murphy, 1968). Substantive goodness captures the idea that forecasters should be able to distinguish between events that occur or not with their confidence estimates (*resolution*). In other words, do confidence ratings reflect whether the judge has made a correct choice or not? The slope (difference between the mean confidence rating for correct choices and the mean confidence rating for incorrect choices) is one measure of substantive goodness (Yates, 1990; Yates & Curley, 1985). As we showed in the present study, the slope scores increased under time pressure in both perceptual and general knowledge tasks (Table 5). This increase in slope is indicative of a resolution in confidence about the accuracy of their decision under time pressure at choice. 2DSD attributes this enhanced resolution to the increase in evidence collection, whether it is with an increase in inter-judgment times τ in the interrogation model or decrease in exit probabilities w in the optional stopping model.

Recall also that this increase in slope is paired with – as 2DSD generally predicts – an increase in scatter or the pooled variance of confidence across correct and incorrect choices (Table 5). This increase in scatter (i.e., greater variance) may detract from the increase in slope (i.e., mean difference) in terms of a judge’s resolution (Wallsten, Budescu, Erev, & Diederich, 1997; Yates & Curley, 1985). To examine this question we calculated a standardized measure of resolution called DI',

\[
DI' = \frac{\text{slope}}{\sqrt{\text{scatter}}} = \frac{\text{conf}_{\text{correct}} - \text{conf}_{\text{incorrect}}}{\sqrt{\frac{n_{\text{correct}} \text{var}(\text{conf}_{\text{correct}}) + n_{\text{incorrect}} \text{var}(\text{conf}_{\text{incorrect}})}{n_{\text{correct}} + n_{\text{incorrect}}}}}
\]
(Wallsten et al., 1997). Table 8 lists the mean and standard deviations of the DI’ scores across participants in each task (individual estimates can calculated using the values in Table 5). Using DI’ as an index of resolution, we still see an increase in resolution during the speed conditions of both tasks. Baranski and Petrusic (1994) report a similar result. This increase in standardized resolution is best understood within the context of the 2DSD interrogation model. Recall in general in DSD standardized accuracy grows as a linear function of the square root of time (e.g., \( \tau \)) (see Equation 5).\(^{12}\) This finding (increased resolution in confidence judgments when facing time pressure at choice but not during confidence) is added as the eighth and final empirical hurdle any model must explain (Table 1). Race models using a one-stage balance of evidence hypothesis do not clear this empirical hurdle.

**Normative goodness.** The second dimension of the accuracy of subjective probabilities is *normative goodness*. Normative goodness addresses the idea that when confidence ratings come in the form of subjective probabilities, we also demand them to adhere to the properties of probabilities. This adherence is because decision makers use subjective probability judgments like the likelihood of rain tomorrow or the probability that a sports team will win, to weight the costs and benefits of different outcomes in making a variety decisions. Thus, confidence ratings when given as subjective probabilities should also be evaluated in terms of their *normative goodness* or how well they meet the demands of probabilities (Winkler & Murphy, 1968). We can further break normative goodness into *coherence* and *correspondence*. The first factor of normative goodness is coherence or the degree to which estimates conform to the mathematical properties of probabilities specified in the Kolmogorov axioms of probability (see for example Rottenstreich & Tversky, 1997; Tversky & Kahneman, 1974; Tversky & Koehler, 1994). For this paper we focused on the second factor of correspondence or the degree of calibration between estimated subjective probabilities and the true probabilities of an event occurring. For example, if a judge says the probability she is correct is 75% then is she actually correct 75% of the time? Note that correspondence in subjective probability estimates implies coherence, but coherence does not imply correspondence. However, correspondence does not necessarily imply good resolution or substantive goodness. For example, a weather forecaster that uses the long run historical relative frequency of rain during a particular month as his/her forecast might be well calibrated, but certainly does not have good resolution.
Participants in our study were generally overconfident in both the line length and city population task. One measure of correspondence is the difference between the average confidence rating and the proportion correct,

\[ \text{bias} = \bar{\text{conf}} - pc \]  

where \( \text{bias} > 0 \) indicates overconfidence.\(^{13}\) That is judges tend to overestimate the likelihood they are correct. Table 8 lists the mean and standard deviation of the bias scores across participants in the speed and accuracy conditions of both tasks (individual estimates can calculated using the values in Table 3). The bias scores show that most participants were on average overconfident in both the line length and general knowledge task. Past results have sometimes found under-confidence in perceptual tasks like the line length task and over-confidence in general knowledge tasks like the city population task (Björkman, Juslin, & Winman, 1993; Dawes, 1980; Keren, 1988; Winman & Juslin, 1993) though not always (Baranski & Petrusic, 1995). This divergence has sometimes been understood as indicating separate and distinct choice/judgment processes for perceptual and conceptual/general knowledge tasks (Juslin & Olsson, 1997; Juslin et al., 1995). We will return to this two vs. one process argument in the discussion, but note that by and large the 2DSD has provided an adequate account of the data in both tasks suggesting perhaps the distinction between perceptual and conceptual tasks is more a difference in information rather than a difference in process.

Table 8 also shows that in general participants were slightly more overconfident when they were under time pressure. These differences, however, are small. Using the statistics in Table 3 one can see that the change in overconfidence arose because, for example, in the line length task their proportion correct fell about 7 percentage points when under time pressure but their confidence only decreased on average about 5 percentage points. A similar pattern emerged in the city population task where accuracy fell about 7 percentage points while confidence only decreased about 4 percentage points.

Reliability diagrams or calibration curves are a useful way to visualize the correspondence between stated confidence ratings and the proportion of correct inferences. Calibration curves plot the relative frequencies of an event occurring (correct choice) for a given respective discrete forecasts (.50, .60, …, 1.00) (Lichtenstein et al., 1982; Murphy & Winkler, 1977). Figure 3 shows the calibration plot for the two different tasks for participant 2. The hard
conditions represent the hardest three conditions of each task and the easy condition represents the easiest three. The plotted lines represent the values calculated using the 2DSD interrogation model with trial variability. The parameters that were used were the best fitting parameters from fitting the model to the distributions of the cognitive performance indices. The figure shows that the 2DSD model does a pretty good job capturing the changes in calibration across the various conditions. The mean absolute deviation between the observed and predicted calibration curve for this participant was .08 in both the line length and city population tasks. The mean absolute deviation across all participants was .09 for both tasks.

To understand why according to 2DSD there is a small effect of time pressure on overconfidence, we can rewrite the expression for the bias score as,

\[
\text{bias} = pc \cdot (\text{slope} - 1) + \text{conf}_{\text{incorrect}}.
\]

Recall \( \text{slope} = \text{conf}_{\text{correct}} - \text{conf}_{\text{incorrect}} \). In this study with the two-choice half scale (e.g., confidence responses > .5) the slope score is bounded to be less than .5, while the mean confidence in incorrects has to be greater than .5. Therefore, everything else remaining equal, judges will move towards more overconfidence (bias > 0) if there is (a) a decrease in the proportion correct \( (pc) \), (b) an increase \( \text{slope} \), and/or (c) increases in \( \text{conf}_{\text{incorrect}} \). According to 2DSD, the speed condition produces a decrease in \( \theta \) which in turn produces a decrease in the proportion correct \( (pc) \); and there was also an increase in slope (due to an increase in \( \tau \)). But, also according to the 2DSD model the change in slope was due to an interaction between the decrease in \( \theta \) and an increase in \( \tau \) so that \( \text{conf}_{\text{incorrect}} \) was substantially lower in the speed conditions (whereas \( \text{conf}_{\text{correct}} \) was only slightly lower). In short, choice time pressure causes an increase in \( pc \cdot (\text{slope} - 1) \) that is offset by a decrease in \( \text{conf}_{\text{incorrect}} \) so that little change in bias results. Taken together 2DSD implies that the increase in the amount of evidence collected during the confidence judgment helps makes the judgment system fairly robust to the influence of time pressure on calibration.

As is evident in the calibration curves in Figure 3 in both tasks there was what has been called the hard/easy effect where judges tend to grow less overconfident as choices get easier and even exhibit underconfidence for the easiest questions (Ferrell & McGoey, 1980; Gigerenzer et al., 1991; Juslin, Winman, & Olsson, 2000; Griffin & Tversky, 1992; Lichtenstein et al., 1982).
In the line length task, collapsing across the easiest three conditions, across participants the bias score went from an average of .13 ($SE = .004; \text{Std}_{\text{betw}} = .09$) in the hardest three conditions to .00 ($SE = .002; \text{Std}_{\text{betw}} = .05$) in the easiest conditions. In the city population task, these numbers were .15 ($\text{Std}_{\text{betw}} = .07; = .005$) and .02 ($\text{Std}_{\text{betw}} = .06; = .004$), respectively. The 2DSD model attributes this hard/easy effect primarily to two factors: (a) a change in the quality of the evidence judges collect to make a decision ($\upsilon$ in the model with trial variability and $\delta$ in the model without trial variability); and (b) judges not adjusting their confidence criteria with the change in the quality of the evidence (for a similar argument see Ferrell & McGoe, 1980; Suantak, Bolger, & Ferrell, 1996).

**Overall accuracy.** Now we return to the main question that permeates all of these analyses: *Why did participants respond to changes in time pressure by increasing the amount of evidence they collect during the second stage of processing?* Intuitively, the reaction of participants seems sensible. When forced to make a choice under time pressure, it seems reasonable to take a bit longer and collect more evidence to assess one’s confidence in that choice. In fact, using 2DSD we can also see that this reaction is at least consistent with an optimal solution where judges seek to minimize decision and inter-judgment time as well as maximize choice and confidence accuracy.

To derive this prediction, it is useful to measure choice and confidence accuracy (or inaccuracy) with the Brier (1950) score

$$Brier = (\text{conf}_i - \text{correct}_i)^2$$

In this equation, $\text{correct}_i$ is equal to 1 if the choice on trial $i$ was correct otherwise 0 and $\text{conf}_i$ was the confidence rating entered in terms of probability of correct (.50, .60, ..., 1.00). In this case the goal of the judge is to produce judgments that minimize the Brier score. The Brier score is a useful overall measure of accuracy for two reasons. One reason is that the Brier score is a strictly proper scoring rule (Aczel & Pfanzagl, 1966; de Finetti, 1962; Murphy, 1973; von Winterfeldt & Edwards, 1986; Yates, 1990). This means that if participants have an explicit goal of minimizing their Brier score they will achieve this minimum if they (a) give an accurate choice and (b) truthfully map their internal subjective belief that they are correct $p^*$ to the closest external probability value available ($\text{conf}$). Recall participants were rewarded in both tasks.
Modeling choice, decision time, and confidence 43

according to a linear transformation of the Brier score (see Equation 15). Thus, it was in participants’ best interests to have a goal of minimizing their Brier score.

A second reason the Brier score is useful is that the mean Brier score can be decomposed in the following manner:

\[
\text{Brier} = VI + \text{Bias}^2 + (VI)(\text{Slope})(\text{Slope} - 2) + \text{Scatter}
\]

(22)

(Yates, 1990; Yates & Curley, 1985) and for a similar decomposition (see Murphy, 1973). In the above equation, \(VI\) or variability index is set equal to \(VI = p(\text{correct})p(\text{incorrect})\). Thus, according to this decomposition, the Brier score integrates, choice accuracy, normative and substantive goodness (cf., Winkler & Murphy, 1968).

Table 8 lists the average Brier score under speed and accuracy instructions for both the line length and city population discrimination task. The standard error for the average Brier score was estimated using bootstrap methods. By and large, participants tended to have worse Brier scores (higher) in the speed conditions. Not listed in the table is the effect of difficulty on the Brier score. Predictably, the Brier score was also influenced by difficulty. In the line length task the average Brier score was .211 (\(SE = .065; \text{Std}_{\text{Btwn}} = .021\)) in the hardest three conditions which was significantly larger than the average Brier score in the easiest three conditions .045 (\(SE = .019; \text{Std}_{\text{Btwn}} = .006\)). In the city population task, the average Brier score was .260 (\(SE = .029; \text{Std}_{\text{Btwn}} = .019\)) in the hardest three conditions which was significantly larger than the average Brier score in the easiest three conditions .142 (\(SE = .023; \text{Std}_{\text{Btwn}} = .033\)).

Using the Brier score decomposition in Equation 22 we can see how changes in different types of accuracy lead to changes in the Brier score. For example, a decrease in choice accuracy will by and large increase the mean Brier score via the variability index. But, an increase in slope will decrease the Brier score (the derivative of Equation 22 with respect to slope is negative).

According to 2DSD when judges increased the amount of evidence collected during the second stage of processing under time pressure at choice they were at least in principle counteracting the negative impact of lower choice accuracy on the Brier score and their final earnings. To investigate this claim we used the 2DSD interrogation model with trial variability to evaluate how the Brier score changes as a function of (a) choice thresholds \(\theta\); and (b) inter-judgment times \(\tau\). Figure 4 plots this relationship for Participant 3 in the city population task. It shows that according to 2DSD the Brier score will decrease when choice thresholds are increased and/or when inter-judgment times \(\tau\) are increased.\(^{14}\)
But of course this change in choice thresholds and inter-judgment times exacts some costs on the judge. In particular, the improvement in overall accuracy comes at the cost of greater processing time. A way to conceptualize these different demands of accuracy and time is with the cost function

\[ \text{Cost} = c_1 \cdot t_D' + c_2 \cdot \tau' + c_3 \cdot \text{Brier} \] (23)

Where \( c_1, c_2, \) and \( c_3 \) specify the cost of the observed decision time \( t_D' \), the observed inter-judgment time \( \tau' \), and the level of a Brier score, respectively.

Using 2DSD one can ask what response parameters for a given stimulus or stimuli minimize the cost function in Equation 23. In particular, for the 2DSD interrogation model what choice thresholds (\( \theta \)), inter-judgment times (\( \tau \)), and confidence criteria, lead to the smallest cost for a given set of costs? A closed form solution does not appear to be available to answer this question. However, the monotonic decreases in the Brier score with choice thresholds and inter-judgment times demonstrated in Figure 4 imply a unique solution can often be found. In particular if (a) judges seek to minimize their Brier score as the cost of decision time increases, and (b) face a constant level of cost for their inter-judgment time then judges should increase their inter-judgment time to minimize the total cost in Equation 23. That is, by taking more time for judgment following a choice under the choice time pressure condition, participants in the line length and city population task acted in a manner consistent with the optimal solution for this task.

**2DSD Optional Stopping Model Account of Inter-judgment Times**

A new challenge for 2DSD. 2DSD posits a simple and straightforward hypothesis: there is post-decision processing of evidence and judges use this extra evidence accumulation to assess their confidence in their choice. The 2DSD interrogation model captures this idea by supposing judges continue accumulating evidence for a fixed amount of time (i.e., inter-judgment time \( \tau \)) after making a choice treating the inter-judgment times as an exogenous parameter of the model. In fact, when we fit the model we used the observed mean inter-judgment time to estimate the inter-judgment time \( \tau \) parameter in the model. While the simplicity of the interrogation model is certainly a strength and as we have shown the model provides a good account of the data, one unexplained result is the negative correlation between inter-judgment times and confidence shown in Table 4. This negative relationship replicates similar results reported by Baranski and
Petrusic (1998) and Petrusic and Baranski (2003) (see also Van Zandt & Maldonado-Molina, 2004). The 2DSD interrogation model (Figure 1) does not explicitly predict such a relationship.

The relationships between interjudgment times and the other performance-relevant variables in Table 4 suggest that interjudgment time is not an exogenous parameter of the judgment process, but rather endogenous to the process. The strongest relationship is the negative relationship between the level of confidence and the interjudgment times. This relationship between confidence and interjudgment time has been interpreted as further evidence of some postdecisional computational processing (Baranski & Petrusic, 1998; Petrusic & Baranski, 2003; Van Zandt & Maldonado-Molina, 2004).

2DSD can account for this relationship between interjudgment time and confidence as well as the other associations detailed in Table 4 by assuming the stopping rule for a confidence judgment is an optional (rather than a fixed) stopping rule. In this case, during the second stage of evidence accumulation some standard internal to the judgment system determines when a judge stops and makes a confidence judgment (much like the choice threshold θ). To formulate this alternative stopping rule, it is useful to consider 2DSD as a Markov chain (e.g., Diederich & Busemeyer, 2003) as shown in Figure 5. The top chain describes a random walk choice process. The circles represent different evidence states ranging from the lower choice threshold −θ to the upper threshold θ. Evidence is accumulated by adjusting the judge’s evidence state up a step (+Δ) with probability p or down a step (−Δ) with probability q, where the probability p is determined by the mean drift rate parameter δ. The evidence states corresponding to the choice thresholds (black circles) denote the typical absorbing barriers in diffusion models (corresponding to the threshold parameter θ). Once the process reaches one of the thresholds at the end of the chain, a choice is made accordingly. Using the Markov chain approximation, and by setting the step size sufficiently small (Δ), we can calculate the relevant distribution statistics including choice probabilities and decision times that closely approximate a continuous time diffusion process (see Appendix C; Diederich & Busemeyer, 2003).

More importantly for our interests, the discrete state space gives another means to conceptualize our two-stage hypothesis. Under this formulation, the confidence stage is modeled as a second Markov chain (see the bottom chain in Figure 4 when Alternative A is chosen). Now during the second stage of processing, we assume markers (κj) are placed along the evidence state space representing the different confidence ratings (j = .50, .60, ..., ,1.00), one for each
rating. For the intermediary confidence ratings (.60, .70, .80, and .90), each time evidence passes
one of these markers there is a probability $w_{\text{conf}}$ that the judge exits and gives the corresponding
confidence rating. The evidence states representing the confidence rating of .50 and 1.00 were
set equal to an absorbing boundary ($w_{.50} = w_{1.00} = 1.0$, thus the black circles shown in the lower
chain in Figure 4). This means that once the process enters one of these extreme states then with
probability 1 the evidence accumulation process ends and the corresponding confidence rating is
given. Using the same Markov chain methods that determine the choice and decision times, the
distribution of confidence ratings and distribution of inter-judgment times can be computed (see
Appendix C).

**Fitting the 2DSD optional stopping model.** Evaluating the 2DSD optional stopping
model is very challenging because it requires simultaneously fitting the entire distribution of
responses across choices, decision times, confidence ratings, and inter-judgment times for the
different conditions of each decision task. The appendix describes the detailed procedures used
to accomplish the fits. In this section we summarize the most important points. To summarize
the procedure, we fit quantiles averaged across individuals, and this fit was done separately for
each of the two conditions of time pressure and decision task conditions. Each time pressure
and decision task condition had approximately 72 free data points, and the 2DSD optional
stopping model had 14 free parameters. The parameter estimates are shown in Table 8.

The most important finding that needs to be explained with this 2DSD optional stopping
model is the relation between confidence ratings and inter-judgment times. Figure 6 displays the
quantiles (.1, .3, .5, .7, and .9) of the distribution of inter-judgment times, as a function of
confidence and accuracy, averaged across all 6 participants, and separately for each decision task
and speed instruction condition. In particular, the triangles display the median inter-judgment
time, which is largest when confidence is lowest and decreases as confidence becomes more
extreme.

Figure 6 also shows that by and large the model gives a good account of the distributions.
In particular, the model gives a good account of the negative relationship between confidence
and inter-judgment times. One property of the model that helps determine this relationship is the
location of the confidence markers. For example, notice that in our data the .50 confidence rating
had the slowest inter-judgment time. To capture this property the best fitting marker location for
the .50 confidence rating $\kappa_{.50}$ was below the starting point of the choice process of $z = 0$ in the
evidence space (see Table 9). However sometimes there is a non-monotonic relationship between confidence and inter-judgment times where the inter-judgment times associated with “Guess” (or .50 in our study) are slightly faster than the intermediary confidence ratings (Baranski and Petrusic, 1998; Petrusic & Baranski, 2003). The 2DSD optional stopping model could account for this by moving the $\kappa_{.50}$ marker up in the evidence space. Doing so would lead to faster inter-judgment times for this confidence rating.

The 2DSD optional stopping model is also revealing as to why confidence changed as a function of time pressure at choice. In particular, to collect more evidence during the second stage they reduced their exit probabilities, $w_{\text{conf}}$. In the line length task they reduced their exit probabilities by almost 60% and 19% in the city population task. The best fitting parameters in Table 9 also suggest that participants also shifted the location of their confidence markers down, but this shift appears to be commensurate to the amount that their choice thresholds also shifted down. Our analysis suggests this shift in criteria does not offset the impact of changes in choice thresholds and the change in exit probabilities in producing changes in the distribution of confidence ratings.

The distribution of inter-judgment times is also telling of some limitations of the 2DSD optional stopping model and 2DSD in general. In particular, one particular aspect where the optional stopping model falls short is in terms of accounting for the more extreme inter-judgment times associated with 1.00 confidence rating. For example, in Figure 6 the model estimated .9 quantile for the 1.00 confidence rating is always more extreme than the observed .9 quantile. One way to handle this is to assume on some trials judges use an alternative method to make a choice and estimate their confidence. For instance on some trials judges may have direct access to the answer (as in the city population task) or they may use cues extrinsic or intrinsic to the task to also infer their confidence. In either case, it might be assumed that with some probability a process akin to 2DSD is used, but on the other trials they use an alternative process and respond with 100% confidence quickly. Such a two-system process is often assumed in the confidence literature (see Busey et al., 2000; Dougherty, 2001; Koriat, 1997; Wallsten et al., 1999).

Discussion

We have recast the standard diffusion model to give a single process account of choice, decision time, and confidence. Perhaps the best summary of 2DSD is provided by the answer to a recent question from a colleague curious whether “choice, response (decision) time, or
Modeling choice, decision time, and confidence 48

certainty, was a purer measure of the judgment process during a two-alternative forced choice task?” 2DSD implies that no one dependent variable is a pure measure of the judgment process. Instead these three measures simply give different views onto the same underlying evidence accumulation process.

In terms of 2DSD, choice and decision time are products of a standard drift diffusion process where judges accumulate evidence to a threshold and make a choice based on the threshold level of evidence they reach. Confidence reflects a kind of addendum to this process where evidence continues to accumulate after a choice. Judges then use the extra accumulated evidence to select a confidence rating. Thus, in terms of the different views the performance measures provide, choice and decision time reflect the quality and quantity of evidence collected up to the choice threshold. Confidence, in comparison, reflects these two facets as well as (a) the quantity of evidence collected after making a choice; and (b) how the total state of evidence at confidence time is mapped to a confidence rating.

Cognitive underpinnings of confidence

Reframing of past hypotheses. Psychologically, 2DSD reframes and integrates two previous hypotheses about confidence. The first hypothesis is Vicker’s (1979) balance of evidence hypothesis. In 2DSD confidence is – as Vicker’s (1979) originally postulated – a function of the balance of evidence in favor of one alternative over the other. A balance of evidence is in fact a natural product of a DSD model where evidence at any point in the process summarizes the information in favor of one alternative relative to the other. In comparison, to get the balance of evidence in race models, evidence on one counter must be compared to the other counter. Indeed as Vickers and Packer (1982) and others have noted this difference between counters is the natural analog to evidence in diffusion models (p. 183).

The second hypothesis 2DSD reframes is Baranski and Petrusic’s (1998; see also Petrusic & Baranski, 2003) post decision processing of confidence hypothesis. This hypothesis in part stemmed from the idea that race models using Vicker’s balance of evidence hypothesis required some post-decisional computation. Baranski and Petrusic (1998), in turn, hypothesized that this computation would require some processing time that cannot be attributed solely to motor time (p. 932). 2DSD posits that the post-decisional processing is not some sort of computation, but rather further collection of evidence from the same distribution of information that helped the judge make a choice. No doubt, between tasks, the sources of the information can vary, for
example, from the perceptual system in the line length task to semantic and/or episodic memory in the city population task. Regardless of the source of the information, we have shown that this continued collection of evidence qualitatively and quantitatively accounts for a large range of empirical hurdles that range from the well-documented negative relationship between confidence and decision time to relatively new phenomena like an increase in the resolution of confidence judgments when participants face time pressure at choice.

**An interaction between choice and confidence stages**

2DSD also reveals that there is an interaction between the choice and confidence stages. When judges were faced with time pressure at choice they lowered their choice thresholds, but they also increased their inter-judgment times. When these two adjustments happen simultaneously then 2DSD predicts the following empirically supported results: (a) an increase in the variance of the confidence ratings (scatter); (b) very little change in mean confidence in correct choices; and (c) a substantial decrease in the mean confidence in incorrect choices. The latter two effects on mean confidence imply that judges’ slope scores (i.e., the difference between mean confidence for correct and incorrect trials) increase under time pressure at choice.

This interaction between choice and confidence stages is difficult for most other models of confidence to explain. Of course the results automatically rule out signal detection models (Green & Swets, 1966), which are silent on the time course of choice and confidence judgments. Other sequential sampling models like the Poisson race model (Merkle & Van Zandt, 2006; Van Zandt, 2000b), the linear ballistic accumulator (Brown & Heathcote, 2008), or the leaky accumulator (Usher & McClelland, 2004) using the balance of evidence hypothesis (based solely on evidence at the time of choice) also cannot easily handle this pattern. As we indicated earlier, race models, however, may be able to account for the interaction between choice and confidence stage if they are given a second stage of evidence accumulation. That is, after making a choice the counters continue racing to a second confidence threshold and confidence is calculated according to the balance of evidence at this second threshold (see for example Van Zandt & Maldonado-Molina, 2004). Future comparisons between two-stage race models and 2DSD will certainly be revealing and better characterize the interaction between the two stages.

This interaction between choice and confidence stages also appears to be difficult for other confidence models to qualitatively handle. Consider for example Ratcliff and Starns (2009) recent model of response time and confidence –RTCON. RTCON is restricted to no-choice tasks
where judges only rate their confidence say from 0% (certain a prespecified item is incorrect) to 100% (certain a pre-specified item is correct). The basic idea of the model is that each confidence rating is represented with an independent diffusion process. So with say 11 confidence ratings there are 11 racing diffusion processes. The first diffusion process to win determines the confidence rating. To find the drift rate for each diffusion process Ratcliff and Starns (2009) assumed that a stimulus item at test produces some degree of activation which is normally distributed. For example, in their study they focused on recognition memory, so the degree of match between a test item and contents in episodic memory determined the level of activation (with old items having higher levels of activation). This distribution of activation is then divided into different regions, one for each confidence rating, and the integrated activation within each region determines drift rate of the corresponding diffusion process. Ratcliff and Starns (2009) showed that the RTCON model can explain several different phenomena related to estimated receiver operator characteristic (ROC) functions from confidence ratings in recognition memory tasks.

In terms of the with-choice tasks studied in this article, where judges first make a choice and then rate their confidence, there appear to be two natural ways to apply the RTCON model. One is to assume confidence and choice are produced simultaneously. That is, one choice (e.g., left) is mapped to a subset of the confidence ratings (0 to 50%), and the other remaining ratings are mapped into the right option. Thus, the winning diffusion process produces a choice and a confidence rating. Obviously, though, this assumption would face the same problems that the race models do in failing to explain how confidence is dependent on the post-decision evidence accumulation. A second possibility is to assume again a two-stage process where first the choice is made according to a standard two-choice diffusion model and then a second process akin to RTCON determines the confidence rating. Without, however, a clear theory describing how the evidence collected in stage 1 influences the drift rates of the racing diffusion processes in the second stage, it is not possible to derive the observed effects of time pressure at choice on the later confidence ratings.

**Possible models of the no-choice task**

Nevertheless Ratcliff and Starns’ (2009) RTCON model does expose a weakness in 2DSD: it is restricted to tasks where judges first make a choice and then state a confidence rating. One solution to this problem is to assume that in no-choice, full confidence scale
procedures, judges implicitly make a choice and then select a confidence rating. In this case, the distribution of confidence ratings is a weighted average of the distribution of ratings from a correct and an incorrect choice. Indeed this hypothesis does not seem implausible. The Poisson race model, for example, makes a similar implicit hypothesis to model these situations (Van Zandt, 2000b). The methods employed in these no-choice, full confidence procedures may even encourage participants to implicitly make a choice. For instance, during these tasks participants are often instructed that responses above a certain rating (e.g., 50%) indicate some level of confidence that a pre-specified alternative is correct (or true or new), and ratings below the same value indicate some level of confidence that a pre-specified alternative is incorrect (see for example Lichtenstein et al., 1982; Van Zandt, 2000b; Wallsten, Budescu, & Zwick, 1993).

A second solution is to adapt the optional stopping assumptions of 2DSD to the no-choice task. In this case, the judgment process is modeled with a one-stage DSD Markov chain with markers laid out across the evidence space representing the different confidence ratings. As in the 2DSD optional stopping model, for the intermediary confidence ratings each time the evidence passes the marker there is a probability $w_j$ that the judge exits and gives the respective confidence ratings. The extreme confidence ratings would, in turn, be modeled as absorbing boundaries so that if the evidence accumulated to one of these markers it is certain that the respective confidence rating would be given. Both of these models – an implicit choice or one-stage DSD with confidence markers – makes precise and testable predictions regarding the distribution of observed response times and confidence ratings.

**Stopping rules of the second stage**

Before proceeding we should address the two different approaches we used to model the stopping rule in the second stage of evidence accumulation in 2DSD. One approach was to use a stopping rule where something external to the judgment system cues the judge to stop collecting evidence and make a confidence judgment. This is the assumption of the 2DSD interrogation model that implies that inter-judgment times are exogeneous to the judgment process. Our data show that (right or wrong) this model using the observed inter-judgment times as estimates of $\tau$ gives an extremely good account of choice, decision time and confidence at the distribution level. These are the primary measures used in cognitive research.

The data, however, also suggest that the inter-judgment times may be determined by the judgment system itself. In other words, inter-judgment times may be an endogenous variable.
The 2DSD optional stopping model captures this idea where during the second stage of evidence accumulation markers are spread throughout the space. These markers are in turn used to select a confidence rating based on the location of the accumulated evidence. This model not only accounts for the 3 primary measures of cognitive performance, but also gives a good account of the distribution of inter-judgment times. By and large, these two models mimic each other in terms of the predictions of choice, decision time, and confidence, and the effects of the different conditions on these variables. Moreover, while both models are mathematically feasible, the 2DSD interrogation model due to its continuous nature is computationally easier to fit. These properties make the 2DSD interrogation model advantageous for a number of applications seeking to understand cognitive performance via measures of choice, decision time, and confidence. Examples of applications include the instances from the decision sciences (e.g. Arkes, 2001), human factors (e.g., Sanders & McCormick, 2002), and any other applied setting where confidence/subjective probabilities are a crucial measure of one’s belief that an event has or will occur (de Finetti, 1962).

In many ways, the distinction between these two stopping rules is reminiscent of the difference between signal detection (Macmillan & Creelman, 2005) and random walk/diffusion models where the former are certainly oversimplified models of the detection process. In the same way, while the 2DSD optional stopping model may more precisely capture the cognitive process, the 2DSD interrogation model still provides a fairly accurate theoretically grounded tool to measure cognitive performance. In short, if one is only interested in choice, decision time, and confidence ratings, then the interrogation model is more practical; but if one is also interested in predicting judgment time, then the optional stopping model is required.

A common choice and judgment process

There have been several empirical studies that have compared the ability of judges to assess their confidence in perceptual and general knowledge or intellectual tasks (Dawes, 1980; Juslin & Olsson, 1997; Juslin et al., 1995; Keren, 1988; Winman & Juslin, 1993). In these studies, a dissociation was reported between these two domains where judges were found to be overconfident in general knowledge tasks, but underconfident in perceptual tasks. This dissociation along with the fact that many participants make the same systematic mistakes in general knowledge tasks have been interpreted as evidence that judges use distinctly different judgment processes in the two tasks (cf. Juslin & Olsson, 1997; Juslin et al., 1995). More
specifically, the hypothesis has been that confidence judgments in the perceptual domain are based on real-time sensory samples as in a sequential sampling model, but confidence in general knowledge tasks is inferred from the cue or cues used in a heuristic inferential process, such as Take the Best (Gigerenzer et al., 1991). This latter inferential process may also be understood as a sequential sampling process (Lee & Cummins, 2004).

In terms of overconfidence and bias, we did not find a dissociation between the two tasks. Instead, by and large participants were overconfident in both the perceptual line-length and general knowledge city-population tasks and their bias decreased as the stimuli got easier. Statistically, one possible explanation for this difference between levels of bias is that judges in our study on average gave higher confidence ratings in the perceptual task (.86 in the speed condition to .91 in the accuracy conditions) than participants in other studies (e.g., .65 to .68 in study 1 in Keren (1988). But, cognitively, we showed that in both tasks over a variety of conditions 2DSD gives a reasonably good account of the distributions and changes in distributions of cognitive performance indices ranging from choice proportions to decision times to confidence ratings to even inter-judgment times. This implies that a single choice and judgment process may underlie both tasks. Whether this process is implemented in the exact same cognitive/neural systems or if two different systems mimic each other in terms of process is a question for future research.

Indeed, arguments for a common decision process are being made in studies of the neural basis of decision making. Provocative results from this area suggest that sequential sampling models like diffusion models are a good representation of the neural mechanisms underlying sensory decisions (Gold & Shadlen, 2007) which appear to be embedded in the sensory-motor circuitry in the brain (Hanes & Schall, 1996; Kim & Shadlen, 1999; Romo, Hernandez, Zainos, Lemus, & Brody, 2002; Shadlen & Newsome, 2001). These results have led to the hypothesis that these sensory-motor areas are the mediating mechanisms for other types of abstract and value-based decisions (Busemeyer, Jessup, Johnson, & Townsend, 2006; Shadlen, Kiani, Hanks, & Churchland, 2008). While our results do not speak to the underlying neural level, they are consistent with this hypothesis that the same choice and judgment process is used to make a range of decisions. The only difference between these domains is the information feeding the decision process. 2DSD, in fact, extends this hypothesis suggesting the same evidence accumulating mechanism(s) may be used to make confidence judgments.
Accuracy of confidence

Understanding the dynamic process underlying choice and confidence judgments has practical and theoretical implications for our understanding of the accuracy of confidence judgments. There have been several descriptive theories as to why, when and how these judgments are accurate or inaccurate ranging from heuristic accounts (Tversky & Kahneman, 1974) to memory accounts (Dougherty, 2001; Koriat, Lichtenstein, & Fischhoff, 1980; Sieck, Merkle, & Van Zandt, 2007) to aspects of the statistical environment (Gigerenzer et al., 1991; Juslin, 1994) to a stochastic account (Budescu, Erev, & Wallsten, 1997; Erev et al., 1994) to a measurement account (Juslin, Winman, & Olsson, 2000). The focus is not without warrant. Many everyday decisions (like whether to wear a rain poncho to work) or many not-so everyday decisions (like whether to launch the space shuttle, see Feynman, 1986) are based on people’s confidence judgments. Time and time pressure, however, is also an important factor in human judgment and decision making (cf. Svenson & Maule, 1993). Yet, few if any of the descriptive and normative theories of the accuracy of subjective probabilities address the effects of time pressure on the accuracy of subjective probabilities.

2DSD shows that the time course of confidence judgments can have pervasive effects on all the dimensions of accuracy from the substantive goodness (resolution) of confidence judgments to the normative goodness (calibration) of these same judgments to the overall accuracy of the choice and judgments. But, more importantly 2DSD reveals how judges can strategically use the time course of the confidence process to their advantage. In particular, judges can increase resolution by increasing the amount of evidence that they collect during the second stage of 2DSD. In fact, this increase in resolution underlies the reason why, according to 2DSD, if judges have the goal to minimize choice and inter-judgment time and maximize choice and confidence accuracy, then when they face time pressure at choice the optimal solution is to increase inter-judgment time. Participants in our study appear to use this tactic.

More generally, this pattern of findings demonstrates that without understanding the goals of judges in our study – and the role of accuracy within these goals – we would not understand the observed behavior of judges. At the same time, though, the increase in inter-judgment time does not make sense unless we understand the process underlying choice and confidence in terms of 2DSD. This addresses a larger issue in judgment and decision making where there has been a call for basic judgment research to orient away from questions of...
response accuracy and instead focus more on the process (Erev et al., 1994; Wallsten, 1996). Our results actually speak to a broader call for theories of judgment and decision making. We can not focus solely on process or accuracy. Rather we must use and understand both process and accuracy (and more generally judges’ goals) in tandem to explain choice and judgment.

**Conclusion**

Vickers (2001) commented that “despite its practical importance and pervasiveness, the variable of confidence seems to have played a Cinderella role in cognitive psychology - relied on for its usefulness, but overlooked as an interesting variable in its own right.” (p. 148). 2DSD helps confidence relinquish this role and reveals that a single dynamic and stochastic cognitive process can give rise to the three most important measures of cognitive performance in the cognitive and decision sciences: choice, decision time, and confidence. While 2DSD gives a parsimonious explanation of a number of past and some new results, it also reveals a number of unanswered questions. For instance, how does the various types of time pressure influence subjective probability forecasts and what are the implications for our everyday and not-so everyday decisions? And what are the neural mechanisms underlying confidence judgments, are they the same as those underlying decision? We think 2DSD provides a useful framework for taking on these larger and more difficult questions.
References

Aczel, J., & Pfanzagl, J. (1966). Remarks on the measurement of subjective probability and


68*(1), 33-45.

Associates.

effects of averaging subjective probability estimates between and within judges. *Journal of
Experimental Psychology: Applied, 6*(2), 130-147.


Hamilton, Ontario.

Psychology, 44*(2), 310-329.

67*, 1-15.


Baranski, J. V., & Petrusic, W. M. (1994). The calibration and resolution of confidence in

*Canadian Journal of Experimental Psychology-Revue Canadienne De Psychologie
Experimentale, 49*(3), 397-407.

Experiments on the time to determine confidence. *Journal of Experimental Psychology:
Human Perception and Performance, 24*(3), 929-945.


Modeling choice, decision time, and confidence 65


Author Note

Timothy J. Pleskac, Department of Psychology, Michigan State University. Jerome R. Busemeyer, Department of Psychological & Brain Sciences, Indiana University.

A National Institute of Mental Health Research Service Award (MH019879) awarded to Indiana University supported both the beginning and finishing of this work. The second author was supported by National Science Foundation under Grant No. 0817965. We thank Jim Townsend, Thomas Wallsten, and Avi Wershbale, for their input on this work. We also thank Andrew Heathcote, Roger Ratcliff, Marius Usher, Trish Van Zandt, E. J. Wagenmakers, one anonymous reviewer, and members of TJP’s 2010 Nature and Practice of Cognitive Science class for comments on the manuscript. We are also appreciative of Kate LaLonde and Kayleigh Vandenbussche for their assistance in data collection. Various components of this paper were presented at the 2007 Annual Meeting for the Cognitive Science Society; the 2007 Annual Meeting for the Society for Mathematical Psychology: the 2008 Annual Meeting for the Society for Mathematical Psychology; the 2008 Annual Meeting for the Society for Judgment and Decision Making; the 2009 Biennial Conference on Subjective Probability, Utility and Decision Making; 2009 Annual Conference of the Psychonomic Society.

Direct correspondence about this article to Timothy J. Pleskac, Department of Psychology, Michigan State University, East Lansing, MI, 48824. Email: tim.pleskac@gmail.com
Appendix A

The Dynamic Signal Detection Model

We list the relevant distribution formulas for a diffusion process below. The derivations have been published elsewhere (see for example Cox & Miller, 1965; Feller, 1968; Luce, 1986; Ratcliff, 1978; Smith 1990; 2000)

If presented with stimulus $S_A$, assuming a drift rate $\delta$, starting point $z$, and choice threshold $\theta$, and drift coefficient $\sigma$, the probability of choosing alternative $A$, $R_A$ for a Wiener process is

$$P(R_A \mid S_A) = \frac{\exp \left( \frac{4 \delta \theta}{\sigma^2} \right) - \exp \left( \frac{2 \delta (\theta - z)}{\sigma^2} \right)}{\exp \left( \frac{4 \delta \theta}{\sigma^2} \right) - 1}. \quad (A1)$$

The probability of incorrectly choosing $R_A$ when presented with $S_B$, can be found by replacing $\delta$ with $-\delta$ in Equation A1. The expressions when $R_B$ is given can be found by replacing $(\theta - z)$ with $(\theta + z)$.

The finishing time pdf for the time that the activation reaches $\theta$ and the judge responds given stimulus is

$$g(t_D \mid R_A, S_A) = \frac{1}{P(R_A \mid S_A)} \frac{2 \theta}{\sigma^2} \exp \left[ \frac{\delta (\theta - z)}{\sigma^2} \right] \sum_{k=1}^{\infty} k \sin \left[ \frac{k \pi (\theta - z)}{2 \theta} \right] \exp \left[ -\frac{t_D}{2} \left( \frac{\delta^2}{\sigma^2} + \left( \frac{\pi k \sigma}{2 \theta} \right)^2 \right) \right] \quad (A2)$$

The cumulative distribution function is

$$G(t_D \mid R_A, S_A) = P(R_A \mid S_A) - \frac{2 \theta}{\sigma^2} \exp \left[ \frac{\delta (\theta - z)}{\sigma^2} \right] \sum_{k=1}^{\infty} \frac{2k \sin \left[ \frac{k \pi (\theta - z)}{2 \theta} \right]}{\frac{\delta^2}{\sigma^2} + \left( \frac{\pi k \sigma}{2 \theta} \right)^2} \quad (A3)$$
The expressions for the pdf and cdf of the finishing times when stimulus $S_B$ is present can be found by replacing $\delta$ with $-\delta$ and exchanging the choice probability. The expressions when $R_B$ is given can be found by replacing $(\theta - z)$ with $(\theta + z)$ and again changing the choice probability.

Several items should be noted here on using equations A1 to A3. In some models using a random walk/diffusion process it proved necessary to not assume the drift rate (or the mean drift rate) for $S_B$ to be the negative of the drift rate of $S_A$, $-\delta$ (see Ratcliff, 1978; Ratcliff & Smith, 2004). A second point is that in using equations A2 and A3 to calculate the PDF and CDF the so-called fly in the ointment is the summation to infinity. To work around this Racliff and Tuerlinckx (2002) recommend iteratively summing the expression within the sum until the current term and previous term are both less than $10^{-29}$ times the current sum (p. 478).

Programming these in Matlab, I used a fixed value of $k$ that always met this requirement (~200). An alternative method is to use a numerical routine like Voss and Voss’s (2008) fast-dm routine. Finally, equations the PDF and CDF in equation A2 and A3 grow unstable as $t_D$ gets very small.

Van Zandt et al. (2000) suggest using alternative forms of the PDF and CDF for very small values.

From Link and Heath (1975), the mean time to choose $R_A$ when presented with $S_A$ is

$$E(t_D | R_A, S_A) = \frac{1}{\delta} \left[ \frac{2\theta \left( \exp \left( -\frac{2(\theta + z)\delta}{\sigma^2} \right) + \exp \left( -\frac{2(\theta - z)\delta}{\sigma^2} \right) \right)}{\exp \left( -\frac{2(\theta + z)\delta}{\sigma^2} \right) - \exp \left( -\frac{2(\theta - z)\delta}{\sigma^2} \right)} - \frac{(\theta + z) \exp \left( -\frac{2(\theta + z)\delta}{\sigma^2} \right) + (\theta + z)}{\exp \left( -\frac{2(\theta + z)\delta}{\sigma^2} \right) - 1} \right].$$

(A4)

The mean time to choose $R_B$ when presented with $S_A$ is

$$E(t_D | R_A, S_A) = \frac{1}{\delta} \left[ \frac{2\theta \left( \exp \left( -\frac{2(\theta + z)\delta}{\sigma^2} \right) + \exp \left( -\frac{2(\theta - z)\delta}{\sigma^2} \right) \right)}{\exp \left( -\frac{2(\theta + z)\delta}{\sigma^2} \right) - \exp \left( -\frac{2(\theta - z)\delta}{\sigma^2} \right)} + \frac{(\theta - z) \exp \left( -\frac{2(\theta - z)\delta}{\sigma^2} \right) + (\theta - z)}{\exp \left( -\frac{2(\theta - z)\delta}{\sigma^2} \right) - 1} \right].$$

(A5)

The distributions for confidence ratings are given in the text Equation 14.

Trial variability in the model parameters was modeled as follows (see Ratcliff, 1978; Ratcliff & Smith, 2004). The value of the drift rate between trials was assumed to be normally
distributed with a mean $\nu$ and a standard deviation $\eta$, $f(\delta) \sim N(\nu, \eta)$. The value of the starting point was assumed to be uniformly distributed with a range $s_z$, $u(z) \sim \text{Uniform}(s_z)$. The choice probabilities, confidence distributions, as well as the marginal pdf and cdf for the finishing times are then found by integrating across all values of $\delta$ and $z$.

Appendix B

**Quantile maximum probability method for fitting decision times and confidence**

After removing contaminant trials, the raw data in the line length task contains approximately 5,615 total trials with observed choices, decision times, confidence, and inter-judgment times per person. In the city population task this number was 3,799. This high number of trials made it possible to fit 2DSD interrogation model to the distributions of choices, decision times, and confidence ratings. In principle, we could fit both the 2DSD model to the multivariate distribution using maximum likelihood methods. However, the density function for decision times in the 2DSD model can be computationally a time consuming calculation. Instead we adapted Heathcote et al.’s (2002) quantile maximum probability (QMP) estimation method to simultaneously fit models to decision time and confidence rating distributions for corrects and incorrects. In the QMP method, decision time distributions are summarized with quantiles. We used .1, .3, .5, .7, and .9. In words, we found the quantiles that corresponded to points in the decision time distributions where 10, 30, … 90% of the decision times fell at or below that point.

The basic idea of the QMP method is that the quantile estimates form six categories of decision times. Within each category we can determine the frequency of decision times falling between the two boundaries (e.g., 20% of the responses fall within the quantiles for .1 and .3). Then using the multinomial distribution function we can calculate the likelihood of the data $L$ (e.g., the likelihood of incorrect decision times during the speeded line discrimination task) for each model using the cumulative distribution function of decision times for a particular model (e.g., the 2DSD model). The likelihood of the data given a particular model for one task (e.g., line discrimination) is then $L = L_{\text{speed, correct}} \times L_{\text{speed, error}} \times L_{\text{accuracy, correct}} \times L_{\text{accuracy, error}}$.

If we expand the calculation to also simultaneously incorporate the distribution of confidence ratings, then the data representation is a 6 (decision time categories) x 6 (confidence ratings) data matrix for corrects and incorrects for both tasks. We fit the 2DSD interrogation model to this 6 x 6 data matrix for corrects and incorrects for both the city and line length tasks.
Notice that the 2DSD model without trial variability predicts that for a given stimulus with a particular drift rate the distribution of decision times are independent from each other conditional on if the choice was correct or not. That is, for correct and incorrect choices for a given stimulus the distribution of decision times and confidence ratings are independent of each other. Thus, fitting the marginal distributions will produce the same result as fitting the joint distribution of decision times and confidence ratings.

In both the line length and city population tasks we broke the data down into two levels of time pressure crossed with different levels of difficulty. Recall, during the line length discrimination task, we collapsed the sixth and fifth levels of difficulty for both speed and accuracy forming five levels of difficulty in the line length task to model. In the city population task, we formed six different levels of difficulty with approximately 300 pairs in each condition.

The 2DSD models were fit at the individual level to a line length discrimination task where there were 10 conditions (speed vs. accuracy X 5 levels of difficulty). Representing the data in terms of our adapted QMP method, each condition had 60 free data points producing a total of $71 \times 10 = 710$ data points per participant. The city task with 12 conditions (speed vs. accuracy X 6 levels of difficulty) had $71 \times 12 = 852$ free data points per participant. We fit highly constrained models to each participant’s data (for a description of the constraints see the model estimation section). In total there were 16 free parameters (for 710 free data points) in the line length task and 17 free parameters (for 852 free data points) in the city population task.

To estimate the maximum likelihood of each of the sequential sampling models in the QML framework we used an iterative Nelder-Mead (Nelder & Mead, 1965) method (cf., Van Zandt, 2000a). During this procedure, the set of parameters were searched that maximized the QMP function using the Nelder-Mead simplex routine (available in Mathwork’s Matlab). As a means of minimizing the risk of finding local maxima, the simplex routine was iterated many times (5 to 10 times). Each iteration used the previously best parameter values perturbed with random error as starting points. We repeated the iterative routine with several different starting points. One starting point used starting values approximated from previous fits in the literature (Ratcliff & Smith, 2004), another used algebraic approximations calculated from the mean decision times and choice probabilities (see Wagenmakers et al., 2007), and finally another iteration used the best fitting average value across participants from the previous attempts.
Appendix C

Derivations of Markov Chain Approximation of 2DSD Model with the Marker Hypothesis of Confidence

This appendix describes the Markov chain approximation of the 2DSD optional stopping model. The choice probabilities, expected decision times, expected distribution of confidence ratings, and expected inter-judgment times are given. The reader is referred to Diederich and Busemeyer (2003) for a more in depth development of the use of Markov chains to approximate random walk/diffusion models (see also Busemeyer & Diederich, 2010 starting on p. 104).

In the model the choice stage works as follows. The state space of evidence L ranges from the lower choice threshold \(-\theta\) to the upper threshold \(\theta\) as a function of step size \(\Delta\). Consequently, L can be expressed as a function of step size

\[
L = \begin{cases} 
-k\Delta, & -\theta \\
-(k-1)\Delta, & \ldots \\
-\Delta, & (m-1)/2 \\
0, & \Delta, \\
\ldots & (m-1) \\
k\Delta & m
\end{cases}
\] 

(C1)

Where \(\theta = k\Delta\) and \(-\theta = -k\Delta\). The transition probabilities for the m states of the Markov chain are arranged in an m x m transition probability matrix P with the elements \(p_{1,1} = 1\) and \(p_{m,m} = 1\), and for \(1 < i < m\),

\[
P_{i,j} = \begin{cases} 
\frac{1}{2\alpha} \left( 1 - \frac{\delta(-k\Delta + (i-1)\Delta)}{\sigma^2} \sqrt{\rho} \right) & \text{if } j - i = -1 \\
\frac{1}{2\alpha} \left( 1 + \frac{\delta(-k\Delta + (i-1)\Delta)}{\sigma^2} \sqrt{\rho} \right) & \text{if } j - i = +1 \\
0 & \text{otherwise}
\end{cases} 
\] 

(C2)

Where the drift and diffusion coefficients are \(\delta\) and \(\sigma^2\), respectively. The parameter \(\rho\) is the time interval that passes with each sampled piece of interval. As \(\rho\) approaches zero and setting \(\Delta = \alpha \sigma \sqrt{\rho}\) the random walk will converge to a Weiner diffusion process which has a continuous
Modeling choice, decision time, and confidence

The parameter $\alpha > 1$ is a parameter that improves the approximation of continuous time process. We set $\alpha = 1.5$ and $\rho = .001$. This Markov chain is also called a birth-death process (see Diederich & Busemeyer, 2003).

The transition probability matrix $P = \begin{bmatrix} p_{i,j} \end{bmatrix}$ is presented in its canonical form:

\[
P = \begin{bmatrix} \begin{pmatrix} P_1 & 0 \\ R & Q \end{pmatrix} \end{pmatrix}
\]

\[
\begin{array}{ccccccc}
1 & | & m & | & 2 & | & 3 & | & \ldots & | & m-2 & | & m-1 \\
1 & | & 1 & | & 0 & | & 0 & | & \ldots & | & 0 & | & 0 \\
M & | & 0 & | & 1 & | & 0 & | & \ldots & | & 0 & | & 0 \\
2 & | & p_{2,1} & | & 0 & | & p_{2,2} & | & p_{2,3} & | & \ldots & | & 0 & | & 0 \\
3 & | & 0 & | & 0 & | & p_{3,2} & | & p_{3,3} & | & \ldots & | & 0 & | & 0 \\
4 & | & 0 & | & 0 & | & 0 & | & p_{4,3} & | & \ldots & | & 0 & | & 0 \\
& | & \vdots & | & \vdots & | & \vdots & | & \vdots & | & \vdots & | & \vdots & | & \vdots \\
m-3 & | & 0 & | & 0 & | & 0 & | & 0 & | & \ldots & | & p_{m-3,m-2} & | & 0 \\
m-2 & | & 0 & | & 0 & | & 0 & | & 0 & | & \ldots & | & p_{m-2,m-2} & | & p_{m-2,m-1} \\
m-1 & | & 0 & | & P_{m-1,m} & | & 0 & | & 0 & | & \ldots & | & p_{m-1,m-2} & | & p_{m-1,m-1} \\
\end{array}
\]

(C3)

With $P_1$ being a $2 \times 2$ matrix with two absorbing states, one for each choice alternative. $Q$ is an $(m-2) \times (m-2)$ matrix that contains the transition probabilities $p_{ij}$ (see Equation C2). $R$ is an $(m-2) \times 2$ matrix that contains the transition probabilities from the transient states to the absorbing states.

With these submatrices the relevant distribution statistics can be calculated (see for example Bhat, 1984). The probability of choosing option $A$, $P(R_A|S_A)$ and the probability of choosing option $B$, $P(R_B|S_B)$) are

\[
\begin{bmatrix} P(R_A|S_A), P(R_B|S_A) \end{bmatrix} = Z \cdot (I - Q)^{-1} \cdot R. \quad \text{(C4)}
\]
Where Z is a m-2 vector denoting the initial starting position of the process. Assuming no bias then $Z_{(m-3)/2+1}=1$ with all other entries set to 0. I is the identity matrix with the same size as Q, $(m-2) \times (m-2)$. The probability distribution function for the first passage time reaching the boundary with $t = n\rho$, $n = 1, 2, \ldots, \infty$ is

$$ Pr(T = t \mid R_A, S_A), Pr(T = t \mid R_B, S_A) = ZQ^{n-1}R. / \left[ Z(I - Q)^{-1} R \right]. $$

(C5)

The cumulative distribution function is

$$ Pr(T \leq t \mid R_A, S_A), Pr(T \leq t \mid R_B, S_A) = Z(I - Q)^{-1} (I - Q)^n. / \left[ Z(I - Q)^{-1} R \right]. $$

(C6)

The mean decision time conditional on each choice is

$$ E(T \mid R_A, S_A), E(T \mid R_B, S_A) = Z(I - Q)^{-2} R. / \left[ Z(I - Q)^{-1} R \right]. $$

(C7)

Again in the 2DSD framework the evidence accumulation process does not stop once a choice is made, but continues. The Markov chain approximation to the diffusion process allows us to reformulate the second stage more along the lines of process that uses an optional stopping rule. This permits the model to not only predict the distribution of confidence ratings, but also the distribution of inter-judgment times.

In general the model assumes that markers $\kappa_i$ are placed along the evidence state space representing the different confidence ratings (.50...1.00), one for each rating. For the intermediary confidence ratings (.60, .70, .80, and .90), each time the judge passes one of these markers there is a probability $w_i$ that the judge exits and gives the corresponding confidence rating. The evidence states representing the confidence rating of .50 and 1.00 were set equal to an absorbing boundary ($w_{.50} = w_{1.00} = 1.0$, thus the black circles shown in the lower chain in Figure 7).

In what follows, we will describe the Markov chain approximation of this second stage assuming the upper boundary $\theta$ for response alternative A was reached in the first stage. The development is analogous if the lower boundary $-\theta$ was reached except the confidence markers are reflected around the starting point. Under this formulation we attach an additional set of states for all other confidence ratings accept the 1.00 confidence rating (it is associated with the
absorbing boundary). So that the modified state space of the second stage assuming the boundary for response alternative A was reached is

\[ L_A = \{-l\Delta, \cdots, -\Delta, 0, \Delta, \cdots, k\Delta\} \cup \{\kappa_{.60}, \kappa_{.70}, \kappa_{.80}, \kappa_{.90}\} \] .

Where in this case \( \kappa_{.50} = -l\Delta \) and \( \kappa_{1.00} = k\Delta \). The transition probability matrix is similar to that given in Equation D3 except for three changes. A new transition matrix \( P^A \) is formed that is \((m+4) \times (m+4)\) where \( m = k\Delta + l\Delta + 1 \). The values in the transition matrix for \( p_{ij} \) are given by Equation D2. The transition matrix has an additional 4 rows and 4 columns to account for the possibility of giving a confidence rating (conf = .60, .70, .80, and .90) associated with one of the four confidence markers \( \kappa_{\text{conf}} \). \( P^A \) in its canonical form is:

\[
P^A = \begin{bmatrix}
P_{1^A} & 0 \\
R^A & Q^A
\end{bmatrix}
\]
In general it takes a similar form as the probability transition matrix $P$ used during the choice stage. The changes are as follows. To allow for the possibility of exiting the evidence accumulation process and giving a confidence rating the evidence state at row $m$ corresponding to the index of $P^A$ associated with the confidence marker $c_{\text{conf}}$ is multiplied by $(1-w_{\text{conf}})$. As $P^A$ shows the last rows contain all zeroes except for the new absorbing states associated with each confidence rating which are set to (see $P_1^A$). The last four columns of $P^A$ contains all zeros except for the row corresponding to the absorption state for each confidence rating and the row corresponding to the confidence marker, $m_{\text{conf}}$, which is set to $w_{\text{conf}}$ (see $R^A$).

Using the submatrices in the transition matrix $P^A$ the distribution of confidence ratings is

$$Z^A \cdot (I^A - Q^A)^{-1} \cdot R^A,$$

where $Z^A$ is a $m-1$ vector denoting the initial starting position of the process, the location of the choice threshold from the choice stage. The functions for expected inter-judgment times can be found in the same way adapting equations C4 to C8.
We modified the QMP method to include the distribution of inter-judgment times. To do so we calculated the .1, .3, .5, .7, and .9 quantiles for the inter-judgment times for correct and incorrect confidence ratings. Because the study was not designed to investigate the distribution of inter-judgment times, in some cases we have a small number of observations per confidence judgment. Therefore, we then found the average quantiles across the six subjects. The same was done for the correct and incorrect decision times. We fit the model using the predicted distributions of decision time for correct and incorrect choices averaged across the distribution of confidence ratings and the distribution of inter-judgment times for each confidence rating associated with a correct and incorrect choice averaged across decision times.

We fit the full 2DSD optional stopping model to each time pressure condition of each decision task collapsing across the different levels of difficulty. This was done for two reasons. First, in order to have enough data to fit the model at the level of the distribution of inter-judgment times we had to collapse across the levels of difficulty. Second, collapsing across difficulty levels also sped up the fitting process as fitting the Markov approximations can sometimes be fairly intensive in terms of memory needs. We fit the full model to each condition, instead of a constrained model where parameters were set equal across conditions, because preliminary fits showed that the fitting routine had difficulty handling these constraints. Nevertheless the parameter estimates in Table 9 suggest they can be set equal across conditions. Each condition had approximately 72 free data points and the 2DSD optional stopping model had 14 free parameters.
Footnotes

1. This assumption that the drift rate changes sign when the stimulus category changes is sometimes relaxed (see Ratcliff, 1978; Ratcliff, 2002; Ratcliff & Smith, 2004).

2. This unbridled growth of accuracy is also seen as unrealistic aspect of random walk/diffusion models. More complex models such as Ratcliff’s (1978) diffusion model with trial-by-trial variability in the drift rate and models with decay in the growth of accumulation of evidence (Ornstein Uhlenbeck models) (Bogacz et al., 2006; Busemeyer & Townsend, 1993) do not have this property.

3. Ratcliff’s (Ratcliff, 1978; Ratcliff & Smith, 2004) diffusion model places the lower threshold $\theta_B$ at the 0 point and places an unbiased starting point at the half-way point between the upper and lower thresholds.

4. The SPRT model could account for a difference in average confidence in corrects and incorrect judgments for the same option if judges are biased where $z \neq 0$. Even so it cannot account for differences in confidence in correct and incorrect choices between two different alternatives (i.e., hits and false alarms), though we are unaware of any empirical study directly testing this specific prediction.

5. Generalizing results from expanded judgment tasks to situations when sampling is internal, like our hypothetical identification task, has been validated in several studies (Vickers, Burt, Smith, & Brown, 1985; Vickers, Smith et al., 1985).

6. Pierce’s model actually predicts a linear relationship between the proportion correct and the mean confidence rating (Equation 9). In fact, the data from both tasks supported this prediction. In the line length task, regressing the average confidence rating across all six subjects onto the average proportion of correct choices accounts for 98% and 99% of the variance in the speed and accuracy conditions, respectively. To evaluate this prediction in the city population task, we first formed six groups of city pairs based the difference in the ordinal ranking of the city pairs in terms of populations (see model fitting section for more details). Using these groups, Pierce’s model accounted for 95% and 90% of the variance in mean confidence rating in the speed and accuracy conditions, respectively.

7. In terms of ANOVAs, for both the line length and city population discrimination task, the interaction between accuracy and time pressure conditions was significant for 5 out of 6
participants. It was not significant for Participant 2 in the line length task who showed no difference in confidence between time pressure conditions (see Table 3) and not significant for Participant 1 in the city population task who had the lowest accuracy in this task (see Table 3).

8. The variance of the scatter statistic was estimated with bootstrap methods.

9. Mathematically, some of the biologically inspired racing accumulator models are even related to random walk/diffusion theory (Bogacz et al., 2006).

10. The basic problem is that while the distribution of decision times becomes more positively skewed as difficulty increases (Ratcliff & Smith, 2004), the Poisson counter model tends to predict more symmetrical distributions. Moreover, in the Poisson race model evidence is accumulated in discrete evidence counts, which in turn means confidence is a discrete variable as well. This causes problems because best fitting parameters often result in confidence having fewer than 6 possible values. Yet, confidence scales can easily have more than 6 values (e.g., when confidence is couched in terms of probabilities).

11. We did not incorporate trial variability into the non-decision time component \( t_{ED} \). Nor did we incorporate so-called contaminant probability parameters that account for the small proportion of trials in which judges have a delay in their decision time. Both of these parameters have been used in other diffusion models (Ratcliff & Tuerlinckx, 2002).

12. Recall this increase in resolution in the speed conditions is evident even with Goodman and Kruskal’s ordinal measure of association \( \Gamma \) (Table 4) (see Nelson, 1984, 1987 for an argument as to the use of \( \Gamma \) as a primary measure of accuracy).

13. Another measure of overconfidence is the “Conf score” (Erev et al., 1994) which is the weighted average difference between the stated confidence rating and the proportion correct across the different confidence ratings excluding the .5 confidence rating. Due to the large sample sizes, the values for Conf score as well as the same statistic including the .5 response are very similar to the bias score statistic and all conclusions stated within the paper are identical. We use the bias score due to its relationship with the Brier score, which we use in the next section.

14. Note this relationship with the Brier score is dependent to some degree on the location of the confidence criteria/markers and drift rate. For example, at extreme levels of drift rate/mean drift rate changes in inter-judgment time and choice thresholds has little effect on the Brier score.
Furthermore, the relationship also breaks down when judges use what appear to be a-typical locations of confidence criteria/markers (e.g., only using the .5, .9, and 1.00 confidence ratings).

15. A similar procedure has been used to model the indifference response in preferential choice using decision field theory (Busemeyer & Goldstein, 1992; Busemeyer & Townsend, 1992; J. G. Johnson & Busemeyer, 2005).

16. Note this is a different approach then in the previous section where we fit the 2DSD interrogation model by constraining a majority of the parameters to be constant across conditions. We chose to fit the full model to each condition to aid the optimization method in finding the best fitting parameters. The parameter estimates we obtained (see Table 8) suggest that some of the model parameters can be held constant across conditions. For example, the drift rate for each time pressure condition in each decision task appears to change very little between conditions.

17. We collapsed across the different levels of difficulty to increase the number of observations per distribution of inter-judgment times. Moreover, we averaged across participants because even with the high number of observations per subject some of the intermediary confidence levels had a low number of observations.

18. Although trial variability in process parameters (e.g., drift rate $\delta$) can be included (see Diederich & Busemeyer, 2003) in this model, we did not incorporate this aspect of the model due to computational limits of fitting the model with the Markov approximation.
Table 1

*The Eight Empirical Hurdles a Model of Cognitive Performance Must Explain*

<table>
<thead>
<tr>
<th>Hurdle</th>
<th>Description</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Speed/Accuracy Tradeoff</td>
<td>Decision time and error rate are negatively related such that the judge can trade accuracy for speed.</td>
<td>(Garrett, 1922; D. M. Johnson, 1939; Pachella, 1974; Schouten &amp; Bekker, 1967; Wickelgren, 1977)</td>
</tr>
<tr>
<td>2 Positive relationship between confidence and stimulus discriminability</td>
<td>Confidence increases monotonically as stimulus discriminability increases.</td>
<td>(Ascher, 1974; Baranski &amp; Petrusic, 1998; Festinger, 1943; Garrett, 1922; D. M. Johnson, 1939; Pierce &amp; Jastrow, 1884; Pierrel &amp; Murray, 1963; Vickers, 1979)</td>
</tr>
<tr>
<td>3 Resolution of confidence</td>
<td>Choice accuracy and confidence are positively related even after controlling for the difficulty of the stimuli</td>
<td>(Ariely et al., 2000; Baranski &amp; Petrusic, 1998; Dougherty, 2001; Garrett, 1922; D. M. Johnson, 1939; Nelson &amp; Narens, 1990; Vickers, 1979)</td>
</tr>
<tr>
<td>4 Negative relationship between confidence and decision time</td>
<td>During optional stopping tasks there is a monotonically decreasing relationship between the decision time and confidence where judges are more confident in fast decisions.</td>
<td>(Baranski &amp; Petrusic, 1998; Festinger, 1943; D. M. Johnson, 1939; Vickers &amp; Packer, 1982)</td>
</tr>
<tr>
<td>5 Positive relationship between confidence and decision time</td>
<td>there is a monotonically increasing relationship between confidence and decision time where participants are on</td>
<td>(Irwin, Smith, &amp; Mayfield, 1956; Vickers &amp; Packer, 1982; Vickers, Smith et al., 1985)</td>
</tr>
</tbody>
</table>
average more confident in conditions when they are take more time to make a choice. This relationship is seen when comparing confidence across different conditions manipulating decision time (e.g., different stop points in an interrogation paradigm or between speed and accuracy conditions in optional stopping tasks).

<table>
<thead>
<tr>
<th></th>
<th>Slow Errors</th>
<th>For difficult conditions, particularly when accuracy is emphasized, mean decision times for incorrect choices are slower than mean decision times for correct choices.</th>
<th>(Luce, 1986; Ratcliff &amp; Rouder, 1998; Swensson, 1972; Townsend &amp; Ashby, 1983; Vickers, 1979)</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>Fast Errors</td>
<td>For easy conditions, particularly when speed is emphasized, mean decision times for incorrect choices are faster than mean decision times for correct choices.</td>
<td>(Ratcliff &amp; Rouder, 1998; Swenson &amp; Edwards, 1971; Townsend &amp; Ashby, 1983)</td>
</tr>
<tr>
<td>8</td>
<td>Increased resolution in confidence with time pressure</td>
<td>When under time pressure at choice, there is an increase in the resolution of confidence judgments.</td>
<td>(Current Paper; Baranski &amp; Petrusic, 1994)</td>
</tr>
</tbody>
</table>
Table 2

Parameters of the 2 Stage Dynamic Signal Detection Interrogation Model of Confidence

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \delta )</td>
<td>Drift rate. Controls the average rate of evidence accumulation across trials and indexes the average strength or quality of the evidence judges are able to accumulate. In fitting the model to the data, the drift rate was made a random variable drawn from a normal distribution with mean ( \nu ) and variance ( \eta^2 ).</td>
</tr>
<tr>
<td>( \sigma^2 )</td>
<td>Drift coefficient. Responsible for the within-trial random fluctuations. It is unidentifiable within a particular condition. In fitting the model, ( \sigma ) is set to .1.</td>
</tr>
<tr>
<td>( \theta_A, \theta_B )</td>
<td>Choice threshold. Determines the quantity of evidence judges accumulate before selecting a choice. Controls the speed-accuracy tradeoff. In fitting the model, we set ( \theta = \theta_A = \theta_B ).</td>
</tr>
<tr>
<td>( z )</td>
<td>Start point. Determine the point in the evidence space where judges begin accumulating evidence. In fitting the model to the data, the starting point was made a random variable drawn from a uniform distribution centered at ( z = 0 ) with a range ( s_z ).</td>
</tr>
<tr>
<td>( t_{ED} )</td>
<td>Mean non-decision time. This parameter accounts for the non-decision time during the task (e.g., motor time). Observed decision time is a function of the non decision time and decision time predicted by the model, ( t_D' = t_D + t_{ED} ).</td>
</tr>
<tr>
<td>( c_{choice,k} )</td>
<td>Confidence criteria. Section the evidence space off to map a confidence rating to the evidence state at the time a confidence rating is made. In general, assuming confidence criteria are symmetrical for an ( R_A ) and ( R_B ) response, there are one less confidence criteria then confidence levels.</td>
</tr>
<tr>
<td>( \tau )</td>
<td>Inter-judgment time. A parameter indexing the time between when a decision is made and a confidence rating is entered.</td>
</tr>
<tr>
<td>( t_{EJ} )</td>
<td>Mean non-judgment time. This parameter accounts for the non-judgment time during the task. Observed inter-judgment time is a function of the non judgment time and inter-judgment time used in the model, ( \tau' = t_{EJ} + \tau ).</td>
</tr>
</tbody>
</table>

**Trial variability parameters**

| \( \nu \) | Mean drift rate across trials. This parameter indexes the mean quality of evidence |
across trials assuming a normal distribution.

$\eta$

*Standard deviation of drift rate across trials.* This parameter indexes the variability of the quality of evidence across trials assuming a normal distribution.

$s_z$

*Range of starting points.* The range of starting points for the uniform distribution. In fitting the model, this parameter was constrained to be no larger than the smallest choice threshold.
### Table 3

**Proportion Correct, Average Decision Time, Average Confidence Rating, and Average Inter-Judgment Time for Each Participant.**

<table>
<thead>
<tr>
<th>Line Length</th>
<th>Par.</th>
<th>Speed</th>
<th>.82*</th>
<th>.80*</th>
<th>.73*</th>
<th>.76*</th>
<th>.83*</th>
<th>.78*</th>
<th>.79*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision Time (s)</td>
<td>Speed</td>
<td>0.54 (0.15)*</td>
<td>0.52 (0.10)*</td>
<td>0.45 (0.10)*</td>
<td>0.54 (0.09)*</td>
<td>0.51 (0.11)*</td>
<td>0.55 (0.09)*</td>
<td>0.52*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Acc.</td>
<td>0.69 (0.33)</td>
<td>0.72 (0.30)</td>
<td>0.87 (0.57)</td>
<td>1.73 (1.47)</td>
<td>0.70 (0.33)</td>
<td>1.56 (1.24)</td>
<td>1.04</td>
<td></td>
</tr>
<tr>
<td>Confidence</td>
<td>Speed</td>
<td>.82 (.22)*</td>
<td>.85 (.16)</td>
<td>.87 (.17)*</td>
<td>.90 (.15)*</td>
<td>.97 (.12)*</td>
<td>.76 (.18)*</td>
<td>.86*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Acc.</td>
<td>.89 (.18)</td>
<td>.85 (.16)</td>
<td>.94 (.12)</td>
<td>.96 (.08)</td>
<td>.99 (.06)</td>
<td>.84 (.16)</td>
<td>.91</td>
<td></td>
</tr>
<tr>
<td>Inter-judgment time (s)</td>
<td>Speed</td>
<td>0.52 (0.43)</td>
<td>0.61 (0.52)</td>
<td>0.58 (0.32)</td>
<td>1.05 (0.82)</td>
<td>0.31 (0.22)</td>
<td>1.02 (0.60)</td>
<td>0.68</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Acc.</td>
<td>0.50 (0.40)</td>
<td>0.52 (0.31)*</td>
<td>0.49 (0.27)*</td>
<td>0.40 (0.24)*</td>
<td>0.26 (0.11)*</td>
<td>0.69 (0.41)*</td>
<td>0.48*</td>
<td></td>
</tr>
<tr>
<td>City Population</td>
<td>Prop. Correct</td>
<td>Speed</td>
<td>.59*</td>
<td>.69*</td>
<td>.67*</td>
<td>.68*</td>
<td>.68*</td>
<td>.66</td>
<td>.66*</td>
</tr>
<tr>
<td></td>
<td>Acc.</td>
<td>.64</td>
<td>.75</td>
<td>.78</td>
<td>.78</td>
<td>.73</td>
<td>.68</td>
<td>.73</td>
<td></td>
</tr>
<tr>
<td>Decision Time (s)</td>
<td>Speed</td>
<td>0.84 (0.26)*</td>
<td>1.08 (0.17)*</td>
<td>0.91 (0.22)*</td>
<td>1.05 (0.17)*</td>
<td>0.96 (0.24)*</td>
<td>1.11 (0.19)*</td>
<td>0.99*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Acc.</td>
<td>1.16 (0.56)</td>
<td>1.57 (0.58)</td>
<td>2.33 (1.52)</td>
<td>2.74 (1.59)</td>
<td>2.58 (1.74)</td>
<td>2.43 (1.16)</td>
<td>2.14</td>
<td></td>
</tr>
<tr>
<td>Confidence</td>
<td>Speed</td>
<td>.58 (.16)*</td>
<td>.78 (.17)*</td>
<td>.81 (.14)*</td>
<td>.83 (.15)*</td>
<td>.83 (.20)*</td>
<td>.75 (.17)*</td>
<td>.76*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Acc.</td>
<td>.60 (.17)</td>
<td>.80 (.17)</td>
<td>.85 (.12)</td>
<td>.89 (.10)</td>
<td>.87 (.17)</td>
<td>.78 (.16)</td>
<td>.80</td>
<td></td>
</tr>
<tr>
<td>Inter-judgment time (s)</td>
<td>Speed</td>
<td>0.58 (0.48)</td>
<td>0.86 (0.42)</td>
<td>0.66 (0.46)</td>
<td>1.27 (0.78)</td>
<td>1.22 (0.80)</td>
<td>1.72 (0.82)</td>
<td>1.05</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Acc.</td>
<td>0.53 (0.39)*</td>
<td>0.57 (0.22)*</td>
<td>0.32 (0.17)*</td>
<td>0.34 (0.14)*</td>
<td>0.65 (0.44)*</td>
<td>1.35 (0.72)*</td>
<td>0.63*</td>
<td></td>
</tr>
</tbody>
</table>

Values in parentheses are standard deviations of the data. * indicates the condition (speed or accuracy) in which a z test revealed the relevant statistic was smaller using an alpha value of .05 (two-tailed). The column at the far right lists the average value of the relevant statistic calculated by weighting each participant’s respective statistic by the inverse of the variance of the individual statistic. Statistical significance for the average participant was determined using the average standard error, assuming random effects.
Table 4
Average (Between Par. Std) Goodman and Kruskal $\Gamma$ Correlation Coefficient Across All 6 Participants for the Line Length and City Population Discrimination Tasks During Accuracy (below diagonal) and Speed (above diagonal).

<table>
<thead>
<tr>
<th>Line Length</th>
<th>Obj. Difference Between Line Lengths</th>
<th>Accuracy</th>
<th>Decision Time</th>
<th>Confidence</th>
<th>Inter-judgment time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Acc \ Speed</td>
<td>-.13 (.04)*</td>
<td>.34 (.10)*</td>
<td>-.15 (.07)*</td>
</tr>
<tr>
<td>Accuracy</td>
<td></td>
<td>.68 (.05)*</td>
<td>Acc \ Speed</td>
<td>.75 (.09)*</td>
<td>-.31 (.10)*</td>
</tr>
<tr>
<td>Decision Time</td>
<td></td>
<td>-.26 (.08)*</td>
<td>- .26 (.10)*</td>
<td>Acc \ Speed</td>
<td>-.16 (.18)*</td>
</tr>
<tr>
<td>Confidence</td>
<td></td>
<td>.40 (.16)*</td>
<td>-.34 (.30)*</td>
<td>Acc \ Speed</td>
<td>-.52 (.25)*</td>
</tr>
<tr>
<td>Inter-judgment time</td>
<td></td>
<td>-.11 (.10)*</td>
<td>-.18 (.12)*</td>
<td>.22 (.11)*</td>
<td>-.47 (.25)*</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>City Population</th>
<th>Obj. Difference Between City Populations</th>
<th>Accuracy</th>
<th>Decision Time</th>
<th>Confidence</th>
<th>Inter-judgment time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Acc \ Speed</td>
<td>-.06 (.05)*</td>
<td>.18 (.04)*</td>
<td>-.07 (.05)*</td>
</tr>
<tr>
<td>Accuracy</td>
<td></td>
<td>.35 (.08)*</td>
<td>Acc \ Speed</td>
<td>.54 (.09)*</td>
<td>-.17 (.12)*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------------</td>
<td>-------------</td>
<td>-------------</td>
<td>----------------------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>Decision Time</td>
<td>-.10 (.04)*</td>
<td>-.17 (.08)*</td>
<td>Acc \ Speed</td>
<td>-.14 (.16)*</td>
<td></td>
</tr>
<tr>
<td>Confidence</td>
<td>.20 (.05)*</td>
<td>.43 (.06)*</td>
<td></td>
<td>.08 (.03)*</td>
<td></td>
</tr>
<tr>
<td>Inter-judgment time</td>
<td>-.03 (.05)</td>
<td>-.07 (.09)</td>
<td>.12 (.05)*</td>
<td>-.16 (.23)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The average Goodman and Kruskal $\Gamma$ correlation coefficients were calculated by weighting each subject’s respective coefficient by the inverse of the variance of the $\Gamma$ assuming random effects. The values in the parentheses are an estimate of the between participant standard deviation. * = $p < .05$ (two-tailed).
### Table 5

**Scores of Slope and Scatter for Each Participant in the Speed and Accuracy Conditions of Each Task.**

<table>
<thead>
<tr>
<th></th>
<th>Par.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>Ave</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Line Length</strong></td>
<td>Slope</td>
<td>Speed</td>
<td>.22 (.01)</td>
<td>.17 (.01)</td>
<td>.21 (.01)</td>
<td>.20 (.01)</td>
<td>.13 (.01)</td>
<td>.21 (.01)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Accuracy</td>
<td>.19 (.01)*</td>
<td>.17 (.01)</td>
<td>.12 (.01)*</td>
<td>.07 (.00)*</td>
<td>.04 (.00)*</td>
<td>.16 (.01)*</td>
</tr>
<tr>
<td></td>
<td>Scatter</td>
<td>Speed</td>
<td>.039 (.001)</td>
<td>.022 (.001)</td>
<td>.019 (.001)</td>
<td>.015 (.001)</td>
<td>.011 (.001)</td>
<td>.025 (.001)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Accuracy</td>
<td>.028 (.001)*</td>
<td>.023 (.001)</td>
<td>.012 (.001)*</td>
<td>.005 (.000)*</td>
<td>.003 (.000)*</td>
<td>.024 (.001)</td>
</tr>
<tr>
<td><strong>City</strong></td>
<td>Slope</td>
<td>Speed</td>
<td>.08 (.01)</td>
<td>.13 (.01)</td>
<td>.14 (.01)</td>
<td>.15 (.01)</td>
<td>.14 (.01)</td>
<td>.10 (.01)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Accuracy</td>
<td>.08 (.01)</td>
<td>.10 (.01)*</td>
<td>.08 (.01)*</td>
<td>.06 (.01)*</td>
<td>.10 (.01)*</td>
<td>.08 (.01)*</td>
</tr>
<tr>
<td></td>
<td>Scatter</td>
<td>Speed</td>
<td>.023 (.001)*</td>
<td>.027 (.001)</td>
<td>.017 (.000)</td>
<td>.017 (.001)</td>
<td>.037 (.001)</td>
<td>.027 (.001)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Accuracy</td>
<td>.027 (.001)</td>
<td>.026 (.001)*</td>
<td>.012 (.000)*</td>
<td>.009 (.000)*</td>
<td>.028 (.001)*</td>
<td>.024 (.001)*</td>
</tr>
</tbody>
</table>

Values in parentheses are standard errors. The average statistics were calculated by weighting each subject’s respective coefficient by the inverse of the variance of the statistic assuming random effects. * indicates the condition (speed or accuracy) in which a z test revealed the relevant statistic was smaller using an alpha value of .05 (two-tailed).
Table 6

Parameter estimates from the 2DSD interrogation model with trial variability in the drift rate and starting point.

<table>
<thead>
<tr>
<th>Par</th>
<th>Line Length</th>
<th>City Population</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 2 3 4 5 6 Ave.</td>
<td>1 2 3 4 5 6 Ave.</td>
</tr>
<tr>
<td>ν₁</td>
<td>0.0521 0.0375 0.0261 0.0322 0.0372 0.0389 0.0373</td>
<td>0.0118 0.0239 0.0205 0.0181 0.0170 0.0158 0.0179</td>
</tr>
<tr>
<td>ν₂</td>
<td>0.1267 0.1032 0.0759 0.0781 0.1112 0.0871 0.0970</td>
<td>0.0234 0.0456 0.0430 0.0435 0.0533 0.0346 0.0406</td>
</tr>
<tr>
<td>ν₃</td>
<td>0.2171 0.1961 0.1418 0.1661 0.1969 0.1544 0.1787</td>
<td>0.0276 0.0739 0.0627 0.0809 0.0697 0.0554 0.0617</td>
</tr>
<tr>
<td>ν₄</td>
<td>0.2762 0.2578 0.1890 0.2308 0.2472 0.2473 0.2414</td>
<td>0.0506 0.0989 0.0757 0.1036 0.0977 0.0856 0.0854</td>
</tr>
<tr>
<td>ν₅</td>
<td>0.3527 0.3719 0.2655 0.3446 0.3189 0.3762 0.3383</td>
<td>0.0565 0.1497 0.0968 0.1477 0.1259 0.1348 0.1186</td>
</tr>
<tr>
<td>ν₆</td>
<td>0.1477 0.1757 0.0876 0.1292 0.0930 0.1583 0.1319</td>
<td>0.0952 0.2264 0.1476 0.2070 0.2168 0.1991 0.1820</td>
</tr>
<tr>
<td>η</td>
<td>0.0634 0.0490 0.0455 0.0463 0.0536 0.0493 0.0512</td>
<td>0.0484 0.1295 0.0662 0.1085 0.1419 0.1532 0.1080</td>
</tr>
<tr>
<td>θ_{speed}</td>
<td>0.0854 0.0770 0.0932 0.1722 0.0797 0.1440 0.1086</td>
<td>0.0679 0.0674 0.0857 0.0699 0.0834 0.0741 0.0747</td>
</tr>
<tr>
<td>θ_{acc}</td>
<td>0.0580 0.0281 0.0110 0.0378 0.0312 0.0000 0.0277</td>
<td>0.0908 0.1141 0.1570 0.2125 0.2013 0.1958 0.1619</td>
</tr>
<tr>
<td>s_z</td>
<td>0.3234 0.3747 0.2920 0.4084 0.3240 0.3966 0.3532</td>
<td>0.0430 0.0391 0.0000 0.0440 0.0411 0.0000 0.0279</td>
</tr>
<tr>
<td>t_{ED}</td>
<td>0.0660 0.0000 0.0000 0.0000 0.0000 0.0000 0.0112</td>
<td>0.4510 0.7847 0.4590 0.7289 0.5579 0.7709 0.6254</td>
</tr>
<tr>
<td>t_{EJ}</td>
<td>0.0750 -0.0143 -0.0713 -0.0956 -0.0193 0.0706 -0.0092</td>
<td>0.1325 0.0210 -0.0199 -0.1075 0.1147 0.0809 0.0370</td>
</tr>
<tr>
<td>c₁</td>
<td>0.0750 -0.0143 -0.0713 -0.0956 -0.0193 0.0706 -0.0092</td>
<td>0.1325 0.0210 -0.0199 -0.1075 0.1147 0.0809 0.0370</td>
</tr>
<tr>
<td></td>
<td>$c_2$</td>
<td>$c_3$</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td></td>
<td>0.1066</td>
<td>0.1157</td>
</tr>
<tr>
<td>$c_2$</td>
<td>0.0543</td>
<td>0.1251</td>
</tr>
<tr>
<td></td>
<td>-0.0024</td>
<td>0.0491</td>
</tr>
<tr>
<td>$c_3$</td>
<td>-0.0068</td>
<td>0.0509</td>
</tr>
<tr>
<td></td>
<td>-0.0193</td>
<td>-0.0193</td>
</tr>
<tr>
<td>$c_4$</td>
<td>0.1501</td>
<td>0.2151</td>
</tr>
<tr>
<td></td>
<td>0.0471</td>
<td>0.0894</td>
</tr>
<tr>
<td>$c_5$</td>
<td>0.1714</td>
<td>0.1851</td>
</tr>
<tr>
<td></td>
<td>0.0967</td>
<td>0.1553</td>
</tr>
<tr>
<td>$c_2$</td>
<td>0.0518</td>
<td>0.1061</td>
</tr>
<tr>
<td></td>
<td>0.0179</td>
<td>0.1370</td>
</tr>
<tr>
<td>$c_3$</td>
<td>0.1570</td>
<td>0.1936</td>
</tr>
<tr>
<td></td>
<td>0.2226</td>
<td>0.3256</td>
</tr>
<tr>
<td>$c_4$</td>
<td>0.1196</td>
<td>0.1838</td>
</tr>
</tbody>
</table>
Table 7

*The Average Goodman and Kruskal $\Gamma$ Between Confidence and Decision Time Holding Difficulty Constant for the Line Length Discrimination Task*

<table>
<thead>
<tr>
<th></th>
<th>Par.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>6</th>
<th>Ave.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speed</td>
<td>Correct</td>
<td>-.25*</td>
<td>-.34*</td>
<td>.01</td>
<td>-.21*</td>
<td>-.29*</td>
<td>-.21*</td>
</tr>
<tr>
<td></td>
<td>Incorrect</td>
<td>-.08</td>
<td>-.27*</td>
<td>.07</td>
<td>.11</td>
<td>-.15*</td>
<td>-.07</td>
</tr>
<tr>
<td>Accuracy</td>
<td>Correct</td>
<td>-.35*</td>
<td>-.33*</td>
<td>-.30*</td>
<td>-.54*</td>
<td>-.50*</td>
<td>-.39*</td>
</tr>
<tr>
<td></td>
<td>Incorrect</td>
<td>-.12</td>
<td>-.25*</td>
<td>.02</td>
<td>-.21</td>
<td>-.30*</td>
<td>-.18*</td>
</tr>
</tbody>
</table>

Participant 5 used primarily the .50, .90 and 1.00 confidence ratings during the line discrimination task and was thus excluded from these calculations. * $p < .05$ (two-tailed)
Table 8

*DI’, Bias, and Brier Scores Across Participants in the Speed and Accuracy Conditions of Each Task.*

<table>
<thead>
<tr>
<th></th>
<th>Line Length</th>
<th></th>
<th>City Population</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SE</td>
<td>Std&lt;sub&gt;btwn&lt;/sub&gt;</td>
<td>Mean</td>
</tr>
<tr>
<td><strong>DI’</strong></td>
<td>Speed</td>
<td>1.32</td>
<td>0.08</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td>Accuracy</td>
<td>0.99*</td>
<td>0.06</td>
<td>0.16</td>
</tr>
<tr>
<td><strong>Bias</strong></td>
<td>Speed</td>
<td>0.08</td>
<td>0.03</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>Accuracy</td>
<td>0.05*</td>
<td>0.02</td>
<td>0.06</td>
</tr>
<tr>
<td><strong>Brier</strong></td>
<td>Speed</td>
<td>0.141</td>
<td>0.005</td>
<td>0.012</td>
</tr>
<tr>
<td></td>
<td>Accuracy</td>
<td>0.113*</td>
<td>0.006</td>
<td>0.014</td>
</tr>
</tbody>
</table>

DI’ and Brier score standard errors were estimated with a bootstrap method. * indicates the condition (speed or accuracy) in which a z test revealed the relevant statistic was smaller using an alpha value of .05 (two-tailed).
Table 9

Parameter estimates for the 2DSD optional stopping model fit to the average choice, decision time, confidence, and inter-judgment time distributions.

<table>
<thead>
<tr>
<th>Decision Task</th>
<th>Time Pressure Condition</th>
<th>$\delta$</th>
<th>$\theta$</th>
<th>$t_{ED}$</th>
<th>$w_{.60}$</th>
<th>$w_{.70}$</th>
<th>$w_{.80}$</th>
<th>$w_{.90}$</th>
<th>$k_{.50}$</th>
<th>$k_{.60}$</th>
<th>$k_{.70}$</th>
<th>$k_{.80}$</th>
<th>$k_{.90}$</th>
<th>$k_{.99}$</th>
<th>$t_{EJ}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Line Speed</td>
<td></td>
<td>0.0724</td>
<td>0.0709</td>
<td>0.1750</td>
<td>0.0154</td>
<td>0.0107</td>
<td>0.0075</td>
<td>0.0092</td>
<td>-0.1000</td>
<td>-0.0273</td>
<td>0.0014</td>
<td>0.0264</td>
<td>0.0480</td>
<td>0.1457</td>
<td>0.0785</td>
</tr>
<tr>
<td>Length Accuracy</td>
<td></td>
<td>0.0776</td>
<td>0.1088</td>
<td>0.1466</td>
<td>0.0319</td>
<td>0.0291</td>
<td>0.0178</td>
<td>0.0280</td>
<td>-0.0870</td>
<td>0.0116</td>
<td>0.0283</td>
<td>0.0463</td>
<td>0.0495</td>
<td>0.1673</td>
<td>0.0936</td>
</tr>
<tr>
<td>City Speed</td>
<td></td>
<td>0.0329</td>
<td>0.0958</td>
<td>0.2821</td>
<td>0.0137</td>
<td>0.0138</td>
<td>0.0251</td>
<td>0.1075</td>
<td>-0.0686</td>
<td>0.0076</td>
<td>0.0119</td>
<td>0.1671</td>
<td>0.1922</td>
<td>0.1968</td>
<td>0.1053</td>
</tr>
<tr>
<td>Population Accuracy</td>
<td></td>
<td>0.0335</td>
<td>0.1409</td>
<td>0.2894</td>
<td>0.0214</td>
<td>0.0204</td>
<td>0.0250</td>
<td>0.1185</td>
<td>-0.0111</td>
<td>0.0704</td>
<td>0.0745</td>
<td>0.1883</td>
<td>0.2168</td>
<td>0.2206</td>
<td>0.1076</td>
</tr>
</tbody>
</table>
Figure 1. A realization of evidence accumulation in the 2DSD interrogation model of confidence. The black jagged line depicts the process and the predicted distribution of confidence ratings when a judge correctly predicts choice alternative A. To produce a confidence estimate the model assumes after a fixed time interval passes (or inter-judgment time $\tau$) more evidence is collected and an estimate (e.g., .50, .60, …, 1.00) is chosen based on the location of the evidence in the state space. The solid black normal curve on the right hand side of the figure is the distribution of evidence at the time of confidence $t_C$ when a judge correctly chooses alternative A. The dashed normal curve is what the distribution would be if a judge would have incorrectly chose alternative B.
Figure 2. The Latency-confidence-choice probability functions for the average participant during the line length (top row) and city population discrimination task (bottom row). The best fitting functions for the 2DSD interrogation model with trial variability in the starting point and drift rate are shown. The circles with dashed lines are the data, and the squares with solid lines are the fits of the 2DSD model. Unshaded markers are the error or incorrect responses. Shaded markers are the correct responses. The error bars represent ± 1.96 s.e.
Figure 3. Empirical and best fitting (model) calibration curves for Participant 2 in the line length (top row) and city population (bottom row) discrimination task. The easy condition is the easiest 3 levels and the hard condition is the hardest 3 levels in the respective tasks. The error bars represent 95% confidence intervals calculated using the standard error of the proportion correct conditional on the confidence rating category.
Figure 4. Predicted Brier Scores for Participant 3 in City Population Task in Difficulty Level 3. The plot was calculated using the 2DSD interrogation model with trial variability in the parameters. The plot illustrates that according to the 2DSD model increases in the choice threshold $\theta$ and inter-judgment time $\tau$ both minimize a person’s Brier score. This implies that the model can be used to find appropriate choice threshold and inter-judgment time settings that produce the fastest total judgment time (choice + confidence) for a given Brier score. The times are predicted observed choice and inter-judgment times as indicated by the prime.
Figure 5. A Markov-chain approximation of the 2DSD optional stopping model. In the model, evidence accumulates over time toward an upper, $\theta$, and lower threshold, $-\theta$. This accumulation is approximated with discrete states in the model using probabilities $p$ and $q$ of moving a step size $\Delta$ to each adjacent state. This process can produce a trajectory such as the jagged line in Figure 1. After making a choice, judges continue accumulating evidence, but are assumed to lay out markers ($\kappa$) across the state space so that if the process crosses through that particular state judges exit with probability $w_j$ and give the corresponding confidence rating. These markers are identified with grey and black dots in the bottom chain. To adequately fit the data, the model assumed that the two extreme confidence ratings (e.g., .50 and 1.00) were associated with an absorbing boundary so that if the process entered its associated state the judge stops accumulating evidence and states the respective level of confidence. If alternative B was chosen a similar chain is used (not shown), which is a reflection of the chain used if alternative A was chosen. The model predicts at the distribution level choice, decision time, confidence, and inter-judgment times, and their inter-relationships (see Appendix C).
Figure 6. Observed and best fitting (model) distribution of inter-judgment times (τ) as a function of confidence and accuracy in the line length and city population task for the average participant. The observed and best fitting distributions are plotted using the .1, .3, .5, .7, and .9 quantiles of the respective distribution. The best-fitting distribution of inter-judgment times were generated from the 2DSD optional stopping model formulated as a Markov chain. As the figures show the inter-judgment times for both correct and incorrect on average grew faster with increasing levels of confidence and the 2DSD model by and large gives a good account of the data.