RUNNING HEAD: MODELING CHOICE, DECISION TIME, AND CONFIDENCE

Two-Stage Dynamic Signal Detection Theory:

A Dynamic and Stochastic 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PLEASE DO NOT CITE WITHOUT PERMISSION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

May 14, 2009

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 model – the two-stage Dynamic Signal Detection (2DSD) model – 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 cognitive performance. Furthermore, the model also accounts for the distributions of each variable in both a perceptual and general knowledge task. Finally, the dynamic nature of the model reveals the moderating effects of time pressure on the accuracy of choice and confidence. More generally, 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

The cognitive and decision sciences each have a vested interest in understanding confidence. On the one hand, the cognitive sciences use confidence to chart the inner workings of the mind. This is true at the lowest levels where, for example, in psychophysics confidence was originally thought to be a window onto Fechner's perceived interval of uncertainty (Pierce, 1877). It is also true at the higher levels. For instance, confidence ratings are used to test and compare different theories of memory (Ratcliff, Gronlund, & Sheu, 1992; Squire, Wixted, & Clark, 2007; Yonelinas, 1994). On the other hand, the decision sciences use confidence 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).

While confidence is clearly an important measure of cognitive performance, our theoretical understanding of it is limited. For instance, despite the implicit assumption in the cognitive sciences that observed choices, decision times, and confidence ratings, tap the same latent process, by and large, most successful cognitive models only account for two of these three primary measures of performance. For example, a signal detection model assumes 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 (Vickers, 1979). The only class of models that have been applied to all three measures are race models like Vickers' (Vickers, 1979) accumulator model, the Poisson race model (Merkle & Van Zandt, 2006; Van Zandt, 2000b; Van Zandt & Maldonado-Molina, 2004), or more recently Ratcliff and Starns (2009) RTCON model. The accumulator and Poisson models, however, are in general less accurate for describing choice and response time as compared to random walk/diffusion models (Ratcliff & Smith, 2004). So this leaves us with a challenge – is it possible to extend the random walk/diffusion class of models to account for confidence? The purpose of this article is to develop 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 better understand the effect of time and time pressure on the accuracy of subjective probabilities, which are typically understood as a special case of confidence ratings (Adams, 1957; Adams & Adams, 1961; Lichtenstein, et al., 1982; Tversky & Kahneman, 1974). Certainly, the accuracy of subjective probabilities has been well studied in the decision sciences (for reviews see Budescu, Erev, Wallsten, & Yates, 1997; Klayman, Soll, Gonzalez-Vallejo, & Barlas, 1999; McClelland & Bolger, 1994). Unfortunately, most formal models of subjective probability estimates are silent in terms of the effect of time and time pressure on the accuracy of these forecasts. As a result little is known as to how or why the accuracy of novice or professional forecasters' 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 address this limitation, we connect our random walk/diffusion model of confidence to the methods used to assess the accuracy of subjective probability estimates.

Our development of the dynamic signal detection model for confidence is structured as follows. As a first step, we review past attempts to model confidence within random walk/diffusion models and examine the empirical phenomena they explain and fail to explain. This review solidifies the critical empirical phenomena summarized in Table 1– many of which were first marshaled out by Vickers (1979) – that any cognitive model of confidence must address (see also Vickers, 2001). Then we will present a random walk/diffusion model of confidence ratings called the two-stage dynamic signal detection (2DSD) model that accounts for these empirical hurdles. Furthermore, we examine new empirical data that qualitatively and quantitatively tests the model's ability to account for confidence under different levels of time pressure in two different decision making tasks. We also compare the 2DSD model with the Poisson race model in their ability to simultaneously account for the choice, decision time, and confidence, data from this study.[1] Finally, the study does highlight some limitations of the 2DSD model, which motivates the development of a more general formulation of the dynamic signal detection model based on a Markov chain approximation of random walk/diffusion models (Diederich & Busemeyer, 2003). In the end we hope to show that the 2DSD model can not only

help chart the inner workings of the mind, but also map the correspondence between our internal subjective beliefs about events occurring and reality.

--

Insert Table 1

--

Diffusion models of two-alternative forced choice tasks

To begin, consider a standard detection task where a judge encounters a stimulus and is asked to choose between two alternatives (A or B). 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. In any of these situations, the judge is presented with a stimulus which could be drawn from one of two classes and the judge is uncertain about the class from which the stimulus came from. Two observable effects can be expected from the judge's decision. First, because in any of these domains the evidence in support of either choice alternative is imperfect and often weak, the judge can make an error (either of omission or commission). Second, the choice between the two alternatives will not occur instantaneously, but will take time to make (*decision time*). We also know that these two aspects (choice accuracy and decision time) are inversely related (D. M. Johnson, 1939; Schouten & Bekker, 1967; Wickelgren, 1977). This inverse relationship produces the speed-accuracy tradeoff where the judge trades 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.

Random walk / diffusion models describe this compromise between speed and accuracy in terms of two aspects: (a) the quality of information; and (b) the quantity of information judges collect to make a decision (Ratcliff & Smith, 2004). This compromise is diagrammed with a realization of a diffusion process for a particular choice trial in Figure 1. The process is based on three general assumptions. The first assumption is that when presented with a stimulus, a judge begins to sequentially accumulate evidence favoring one alternative over the other. The evidence may come from current sensory inputs and/or past experiences stored in memory (Gold & Shadlen, 2007; Heekeren, Marrett, & Ungerleider, 2008). The second assumption is that the sampled evidence at each time step is subject to random fluctuations. The jagged line shown in Figure 1 illustrates the accumulation of noisy evidence. 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. Diffusion models can be considered a special case of random walk models when the steps become arbitrarily small, and so for the rest of the paper we will refer to this larger class of random walk and diffusion models simply as diffusion models.

Insert Figure 1

The third assumption is that when judges reach a preset level of evidence favoring one alternative over the other, they stop collecting evidence and make a decision accordingly. The horizontal lines labeled with θ_A and θ_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 accumulation process reaches the threshold the evidence 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 the model can predict decision times or the time it takes a judge to reach θ_A or - θ_B (the first passage time). With these assumptions the diffusion model accounts for the speed-accuracy tradeoff (hurdle 1). Increasing the magnitude of θ_i will increase the amount evidence needed to reach a choice. This reduces the impact random fluctuations in evidence will have on choice and as a result increase choice accuracy. Larger θ_i, however, also imply more time will be needed before sufficient evidence is collected. In comparison, decreasing the thresholds θ_i lead to faster responses but also more errors. This ability of the threshold in the diffusion model to control decision times and error rates also implies the model can be used to find the optimal threshold that would give the fastest decision for a given level of accuracy (Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006; Edwards, 1965; Wald & Wolfowitz, 1948).

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).

Often, though, experimental psychologists desire an alternative measure of cognitive performance in the form of confidence. Confidence is easily obtained with a simple adjustment in the empirical procedure: ask judges, after making their choice, 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). 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 this paper, we focus on the more common forced-choice task (Baranski & Petrusic, 1998; Wallsten, 1996) and in the discussion address how our model might be adapted to account for the "no-choice" procedures. The reader is also referred to Ratcliff & Starns (2009) for an alternative sequential sampling account of the no-choice confidence task.

The use of confidence begs the question how diffusion models would explain and predict confidence ratings. Until now there were two ways to derive confidence from diffusion models. Both are limited in terms of the empirical phenomena they explain. Next we will review these models as well as the empirical findings they explain and fail to explain. This set of results is an important set of empirical hurdles that any model of confidence ratings must overcome.

General framework of random walk / diffusion models

To facilitate our review of these models, we first introduce a general notational framework for diffusion models adapted from Ashby (1983) and Townsend and Ashby (1983). Formally, a diffusion model assumes that after stimulus S_i ($i = A, B$) is presented, for each time interval Δt that passes, the judge considers a piece of information $y(t)$ where t denotes a particular time point. After a time length of $t = n(\Delta t)$ the judge will have generated a set of n pieces of information drawn from some distribution $f_i[y(t)]$ characterizing the stimulus. In the most general form of diffusion models, 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,

$$L(t + \Delta t) = L(t) + x(t + \Delta t). \quad (1)$$

Figure 1 depicts a realization of this process for one hypothetical trial where the diffusion process is drifting toward one of two boundaries or thresholds. The upper threshold θ_A represents the preset levels of evidence required to choose alternative A so that once the accumulated evidence crosses this thresholds, $L(t) > \theta_A$, alternative A is chosen. Alternative B is chosen if and when the process exceeds the lower $L(t) < -\theta_B$. The time it takes for the evidence to reach either threshold is the predicted decision time, t_D. 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.[2]

With this formal framework, we are now equipped to examine possible hypotheses for obtaining confidence ratings with diffusion models. The first model we will consider – the Sequential Probability Ratio Test (SPRT) model – is perhaps the most straightforward method of obtaining confidence ratings, but as we show it makes an inaccurate prediction: judges should have the same level of confidence in all choices made under the same levels of time pressure.

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 diffusion model,

$$x(t) = h[y(t)] = ln\left[\frac{f_A(y(t))}{f_B(y(t))}\right]. \quad \mathbf{(2)}$$

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

$$L(t + \Delta t) = L(t) + ln\left[\frac{f_A(y(t+\Delta t))}{f_B(y(t+\Delta t))}\right]. \quad (3)$$

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

$$ln\left[\frac{p(H_A|D)}{p(H_B|D)}\right] = \sum_t ln\left[\frac{f_A(y(t))}{f_B(y(t))}\right] + ln\left[\frac{p(H_A)}{p(H_B)}\right]. \quad (4)$$

Consistent with the diffusion model framework, 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 the fastest decision time and the lowest error rate for the given decision time (Bogacz et al., 2006; Edwards, 1965; Wald, 1947). Therefore, reaching a choice threshold (either θ_A or θ_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 descriptive validity. Stone (1960) and Edwards (1965) used the SPRT diffusion model to model human choice and decision times. 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 internal posterior belief (Equation 4) with a logistic transform to a subjective probability of being correct. However, this rule (or any related monotonic transformation) implies that confidence is completely determined by the threshold values (θ_A or θ_B) which in turn implies that when the choice thresholds remains fixed across trials 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 difficulty of the stimuli. That is, confidence in any particular choice alternative is related to objective measures of difficulty for discriminating between stimuli (Baranski & Petrusic, 1998; Festinger, 1943; Garrett, 1922; D. M. Johnson, 1939; Pierce & Jastrow, 1884; Pierrel & Murray, 1963; Vickers, 1979). This monotonic relationship between stimulus difficulty and observed confidence is hurdle number 2 in Table 1.

A further difficulty for the SPRT account of confidence is that the resolution of the expressed confidence is usually good. In other words, judges' confidence ratings discriminate between correct and incorrect responses (Ariely et al., 2000; Baranski & Petrusic, 1998; Dougherty, 2001; Garrett, 1922; D. M. Johnson, 1939; Nelson & Narens, 1990; Vickers, 1979) even when holding stimulus difficulty 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 with fixed thresholds, however, predicts equal confidence for correctly or incorrectly choosing an option.

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). The next step in trying to model confidence ratings with diffusion theory is to relax the assumption that the accumulated evidence is a likelihood ratio. This relaxation is often done (see for example Ashby, 1983; Link & Heath, 1975; Ratcliff, 1978; Townsend & Ashby, 1983; Wagenmakers, van der Maas, & Grasman, 2007) resulting in a more general framework, which in turn allows us to consider alternative mechanisms generating confidence ratings.

Dynamic Signal Detection Theory

The name of this model 'dynamic signal detection theory' (DSD) – encapsulates the idea that one can understand diffusion models as a logical extension of signal detection theory (Busemeyer & Diederich, 2009; Link & Heath, 1975; Ratcliff & Rouder, 2000; Smith, 2000; Wagenmakers et al., 2007). When viewed through the theoretical lens of diffusion theory, the decision process in signal detection theory uses a fixed sample size of evidence to elicit a decision (cf. Pike, 1973). Diffusion models, in comparison, drop this assumption. In fact, if we consider the speed-accuracy tradeoff, then we see that that the discriminability parameter in signal detection theory (d') is confounded with processing time. That is, although the quality of evidence has not changed when people take longer to make a decision they are typically more accurate. If one uses the standard method for estimating d' (see for example Macmillan & Creelman, 2005), then signal detection theory would incorrectly interpret this increase in accuracy as an increase in psychological discriminability and not a decision criterion factor.

In this DSD model, when a stimulus is present, the observed information – again coming from either sensory inputs or memory retrieval – is transformed into evidence with a well behaved function, $x(t) = h[\, y(t)\,]$, but not necessarily a log likelihood. This transformation allows DSD to encapsulate different processing assumptions including the possibility that the evidence is based on a comparison between the sampled information and a mental standard (Link & Heath, 1975), or a measure of strength based on a match between a memory probe and memory traces stored in long term memory (Ratcliff, 1978). Regardless of the specific cognitive

underpinnings, when stimulus S_A is present 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 present the mean is equal to E [$x(t)$] = $-\delta\Delta t$ and variance V [$x(t)$] = $\sigma^2\Delta t$. Using Equation 1 this diffusion model can be written as a stochastic linear difference equation:

$$dL(t) = L(t + \Delta t) - L(t) = X(t + \Delta t) = \delta h + \sqrt{\Delta t}\varepsilon(t + \Delta t), \quad (5)$$

Where ε(t) is a white noise process with a mean of 0 and variance σ^2. A standard Wiener diffusion model – where evidence accrues continuously over time – is derived when the time step Δt approaches zero so that the above discrete process converges to a continuous time process (Cox & Miller, 1965; Smith, 2000). A consequence of Δt approaching zero is that via the central limit theorem our uncertainty regarding 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 and the diffusion process is unbounded the mean evidence state increases linearly with time

$$E[L(t)] = \mu(t) = n \cdot \Delta t \cdot \delta = t \cdot \delta, \quad (6)$$

and so does the variance

$$V[L(t)] = \sigma^2(t) = n \cdot \Delta t \cdot \sigma^2 = t \cdot \sigma^2 \quad \textbf{(7)}$$

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

$$d'(t) = \frac{2 \cdot \mu(t)}{\sigma(t)} = 2\left(\frac{2 \cdot \delta}{\sigma}\right)\sqrt{t} = 2d\sqrt{t}. \quad (8)$$

Where $d = 2\delta/\sigma$. Recall the evidence for stimulus S_A has a mean of $\delta\Delta t$ and S_B has a mean of $-\delta\Delta t$. In words, Equation 8 states that accuracy grows as a square root of time so that the longer people take to make process the stimuli the more accurate they become.[4] Equation 8 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 *d* or the rate of evidence accumulation 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 our 2DSD model.

The choice threshold (θ_A and θ_B) and bias (z) parameters permit the model to predict choices and with the fully specified DSD model we can derive the expressions for the cognitive performance variables of choice probabilities and mean decision times (see Cox & Miller, 1965 , p. 58). The probability of response R_A when stimulus S_A is present (hit) is

$$P(R_A|S_A) = \frac{exp\left(4\frac{\delta\theta}{\sigma^2}\right) - exp\left(2\left(\frac{\delta(\theta - z)}{\sigma^2}\right)\right)}{exp\left(4\frac{\delta\theta}{\sigma^2}\right) - 1}. \quad (9)$$

The probability of response R_A when stimulus S_B is present $P(R_A|S_B)$ (false alarm) is found by making the drift rate negative, $-\delta$, in Equation 9. The mean time to make a decision when stimulus S_A is present can also be found

$$E(t_D|S_A) = 2k\frac{\left[P(R_A|S_A)\frac{(\theta - z)}{\sigma} - P(R_B|S_A)\frac{(\theta + z)}{\sigma}\right]}{\delta/\sigma}. \quad (10)$$

The parameter k is a scaling factor that accommodates the different time units judges have for collecting evidence. For the current paper we will set k equal to 1. The expressions for the distributions of decision times or finishing times are given in Appendix A.

Pierce's Law

Thus, two of the most important indices of cognitive performance are formally captured with DSD. The model still lacks the ability to model confidence ratings. Based on the previous section, we know that the total amount of evidence judges aim to collect (θ) cannot account for confidence ratings. An alternative account has its roots in perhaps the first formal hypothesis of confidence (cf., Link, 1992, 2003; Vickers, 2001). Pierce (1877) hypothesized that confidence reflected Fechner's perceived interval of uncertainty and as a result speculated that confidence should be logarithmically related to the chance of correctly detecting a difference between stimuli.[5] More formally, Pierce and Jastrow (1884) showed that the average confidence rating in a discrimination task was well described by the expression:

$$\overline{conf} = \beta \cdot ln\left[\frac{P(R_A|S_A)}{P(R_B|S_A)}\right]. \quad (11)$$

The parameter β is a scaling parameter. While innovative and thought provoking for its time, the law is descriptive at best. Link (2003), however, provided a theoretical justification for the function based on the parameters of a diffusion model. To see how, notice that if we assume no bias on the part of the judge in our hypothetical two alternative forced choice task ($z = 0$ and/or $\theta_A = \theta_B$), we can use the correct and incorrect choice proportions of the DSD model,

$P(R_A|S_A)$ *and* $P(R_B|S_A)$, to derive the following relationship between the choice threshold (θ) and drift rate (δ):

$$\overline{conf} = ln\left[\frac{P(R_A|S_A)}{P(R_B|S_A)}\right]/2 = \delta\,\theta/\sigma^2. \quad (12)$$

In words, under unbiased conditions Pierce's Law can be understood as a function of the quantity of the information needed to make a decision (θ; or the distance traveled by the diffusion process) and the quality of the information (δ; or the rate of evidence accumulation in the diffusion process) accumulated in the DSD model (for a more general derivation allowing for bias see Heath, 1984). We will call this function *Pierce's function*.

This hypothesized relationship between drift rate and choice thresholds, on the one hand, and confidence, on the other, surmounts several of the empirical hurdles any model of confidence must pass. Of course the model clears hurdle 1: the speed/accuracy tradeoff. Pierce's function also accounts for the monotonic relationship between difficulty and confidence (hurdle 2) because as countless studies have shown the drift rate systematically decreases with increasing stimulus discrimination difficulty (e.g., Ratcliff & Rouder, 1998; Ratcliff & Smith, 2004; Ratcliff, Van Zandt, & McKoon, 1999). This decrease in drift rate as difficulty increases implies – via Pierce's function – that confidence also decreases.

Notice, though, that Pierce's function is silent in terms of hurdle 3 where confidence in correct choices is greater than in incorrect choices. This is because Pierce' function uses correct and incorrect choice proportions to predict the mean confidence. More broadly *any hypothesis* that posits confidence is a direct function of the diffusion model parameters (δ, θ, 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 function does bring to light two additional hurdles that a confidence model must surmount. Using Pierce's function, we see that a DSD model predicts that there is a negative relationship between decision time and the degree of confidence expressed in the choice. This is because, according to the model, as drift rate decreases (stimulus discrimination becomes more difficult) the average decision time increases while according to Equation 12 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 (Baranski & Petrusic, 1998; Festinger, 1943; D. M. Johnson, 1939; Vickers & Packer, 1982). This negative relationship between decision time and confidence

in *optional stopping tasks* (where observers control their own sampling by choosing when they are ready to make a choice) serves as empirical hurdle 4 (Table 1).

This intuitive negative relationship between confidence and decision time has been the bedrock of several 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 evident in a variety of conditions (hurdle 5) where the longer judges takes to make a decision the higher their confidence. This monotonically increasing relationship between decision time and confidence has been shown in both optional stopping paradigms and *interrogation paradigms* (where an external observer interrupts judges at different sample sizes and asks them for a choice).

For example, in interrogation paradigms using an expanded judgment task (which externalizes the sequential sampling process asking people to physically sample observations from a distribution and then make a choice), Irwin, Smith, and Mayfield (1956) showed that as judges were required to take more observations (hence greater choice thresholds and longer decision processing times) their confidence in their choices increased (for a replication of the effect see Vickers, Smith, Burt, & Brown, 1985).[6] Similarly, if we compare confidence across conditions of optional stopping tasks under different levels of time pressure, 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 (Baranski & Petrusic, 1998; Festinger, 1943; Garrett, 1922; D. M. Johnson, 1939). This set of contradictory findings is an important limitation to Pierce's function, which we will return to shortly. Regardless of the task, the monotonically increasing relationship between decision time and confidence is credited to changes in the quantity of information collected or the threshold parameter (θ) where larger thresholds lead to longer decision times (Equation 10) and greater levels of confidence (Equation 12).

In summary, although Pierce's function appears to have a number of positive traits, it has limitations that make it less than desirable in terms of a diffusion account of confidence. The limitations by and large can be attributed to the fact that confidence in Pierce's function is a direct function of the quantity (θ) and quality (δ) of the evidence used in the DSD model. On the surface, this assumption seems implausible because it implies a judge would have direct

cognitive access to such information. This access leads one to wonder, "If the judge knew δ why is he or she even uncertain?" But, even if this plausibility criticism is rectified in some manner, there is another more serious problem: Pierce's function as well as any other model that assumes confidence is some function of the quantity (θ) and quality (δ) of the evidence cannot clear hurdle 2 that judges are more confident in correct trials than incorrect trials.

An alternative hypothesis is that at the time a judge enters a confidence rating, judges do not have direct access to the quantity (θ) and quality (δ) of information, but instead have indirect access via some form of the actual evidence they accumulated. Next we will present such an account that draws inspiration from signal detection theory's method of accounting for confidence ratings.

The 2 Stage Dynamic Signal Detection (2DSD) Model

As Figure 1 shows, typically when diffusion models are used in psychology they are instantiated with two absorbing boundaries or thresholds (though see Gomez, Perea, & Ratcliff, 2007 for applications of one absorbing boundary). The absorbing boundaries insure that when the accumulated evidence reaches the evidence states of θ_A or θ_B the process ends and it is assumed that the state of evidence remains in that state thereafter (Cox & Miller, 1965). As an alternative account we assume that after making a choice, judges continue to accumulate evidence. Eventually they interrupt the evidence accumulation process to categorize the accumulated evidence into a confidence response category (see Figure 2). That is, our hypothesis is that a judge does not simply shut down his or her evidence accumulation process after making a choice, but continues to think about the two options and accumulates evidence to make a confidence rating. Thus, the confidence rating is a function of the evidence collected at the time of the choice and plus the evidence collected after making a choice.[7]

Figure 2

This idea that judges continue to sample evidence seems psychologically plausible. Empirically, judges appear to continue to collect evidence even after stimuli are hidden from view (or masked) at the same rate as when the objects were seen (Ratcliff & Rouder, 2000).This implies that the decision system is a fairly robust system in which external constraints cannot stop it from accumulating information. Thus, we argue that if external constraints (like removing

a stimulus) do not necessarily stop the evidence accumulation process, then it seems reasonable to assume that internal demands like making a choice will not stop the process either. Furthermore, 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. In fact, methodologically some psychologists adjust their methods of collecting confidence responses 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, …, 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.

The 2DSD describes this choice and confidence progression as occurring in two stages. During the first stage, the choice stage, the same diffusion process occurs as described earlier. Now, though, after making a decision we assume judges continue accumulating a fixed sample of evidence to make a confidence rating (later in the discussion how to relax this fixed sample assumption). To model this, we allow the diffusion process to first reach a threshold, and then, the process continues for a fixed period of time τ or *inter-judgment time*. In most situations, the parameter τ 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 τ is between 500 to 650 ms and appears constant (especially after a number practice sessions) across confidence ratings; (b) if speed is stressed, then τ was higher (~700 to 900 ms) and seemed to vary across confidence ratings. This last property of varying across confidence ratings suggests the inter-judgment time might be a result of the judgment process, but for the time being, we will assume that τ is an exogenous parameter in the model. Again, later we will propose a more detailed model that treats the confidence responses more like an optional stopping process (thus the inter-judgment time becomes an endogenous parameter to the model) and we show some initial results that suggest this alternative framing can account for the changes in inter-judgment times and their distribution.

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 $\tau = n\Delta t$·

$$L(t_C) = L(t_D) + \sum_{i=1}^{n\cdot\Delta t} x(t_D + i \cdot \Delta t). \quad (13)$$

As Figure 2 depicts, analogous to signal detection theory (e.g., Macmillan & Creelman, 2005), judges scale the accumulated evidence $L(t_C)$ onto the possible response categories. In the case of our general detection task there are six levels of confidence (*conf* = .50, .60, …, and 1.00) conditioned on the choice R_A or R_B, $conf_j|R_i$ where j = 0, 1,…, 5. So each judge needs five response criteria for each option, c_{k,R_A} where k = 1, 2, …, 5 to select among the responses. The response criteria, just like the choice thresholds, are set at specific values of evidence. The location of the criteria depend, as in signal detection theory, on the biases of judges and may also be sensitive to the same experimental manipulations that change the location of the drift 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, $c_{k,R_B} = -c_{k,R_A}$). Future research will certainly be needed to identify if and how these confidence criteria move in response to different conditions. With these assumptions, if the judges choose the R_A option and the cumulated evidence is less then $c_{1,R_A} < L(t_C)$ 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 a function of the distribution of possible evidence accumulations at time point t_C. However, the properties of the distributions reflect the fact that we know what state the evidence was in at the time of choice, either θ_A or θ_B. So our uncertainty about its location at t_C is only a function of the evidence accumulated during the confidence period of time τ. These observations allow us to identify the properties of the evidence at t_C at the distribution level conditional on the four different types of responses judges could generate. When stimulus S_A is present and response R_A is chosen the distribution of evidence at time t_C is normally distributed with a mean of

$$E[L(t_C)|\mathrm{S_A}] = \begin{cases} \tau\delta + \theta_A, & if\ R_A\ was\ chosen \\ \tau\delta - \theta_B & if\ R_B\ was\ chosen \end{cases}. \quad (14)$$

The means for stimulus S_B trials can be found by replacing the δ's with $-\delta$. The variance in all cases is

$$\boldsymbol{var}[\boldsymbol{L}(\boldsymbol{t_C})] = \sigma^2\tau. \quad (15)$$

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

$$Pr\left(conf_j \middle| R_A, S_A\right) = P\left(c_{j,R_A} < L(t_C) < c_{j+1,R_A} \middle| \delta, \sigma^2, \theta_A, \tau\right) \quad (16)$$

where c_{0,R_A} is equal to $-\infty$ and c_{6,R_A} is equal to ∞. Similar expressions can be formulated for the other choices. The precise values of $Pr\left(conf_j \middle| R_A, S_A\right)$ 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. For now we will leave the model as specified and examine how it accounts for the empirical hurdles laid out in front of it.

Table 2

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 14). Thus, the model can make similar predictions as Pierce's function (Equation 12), though Pierce's function posits a multiplicative relationship as opposed to an additive one. This relationship between confidence and the choice threshold θ also points out an important observation that we will return to later: the choice and confidence stages are not independent of each other in the model.

The similarity between Pierce's function and the 2DSD model implies that the model can surmount the empirical hurdles we have laid out in much the same way as Pierce's function overcame them. The model still accounts for the speed-accuracy tradeoff (hurdle 1) because we use the standard diffusion model to make the choices. It also uses the fact that as stimulus difficulty increases so does the drift rate (δ) to explain why confidence is monotonically related to stimulus difficulty (hurdle 2). The model also can 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 $\tau\delta + \theta_A$ for hits (response R_A is correctly chosen when stimulus S_A was shown) and $-\tau\delta + \theta_A$ for false alarms (response R_A is correctly chosen when stimulus S_B was shown) (see Equation 14). 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 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 θ during accuracy conditions than in speed conditions. As a result this will move the means of the confidence distribution 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 6) and thus confidence will be greater when judges are forced to take more time to make a choice.

Finally, the 2DSD model also makes explicit why initial 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). 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). These findings highlight an important component any model of confidence should have: a response mapping process. The 2DSD model 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. Thus, one way to understand the conflicting effects of time pressure on confidence is that when the difficulty of the task changes via time pressure and as a result of the instructions to spread their confidence ratings out across a session, participants change their confidence criteria.

An additional advantage of making the confidence mapping process explicit in the 2DSD model is that it does not restrict the model to a specific scale of confidence ratings. The 2DSD model 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 model accounts for a wide range of historical data sets (hurdles 1 through 5 in Table 1), which also rules out a number of possible alternative theories of confidence rooted within diffusion theory as well as many other theories (Vickers, 1979). However, this is only part of our charge in developing a diffusion theory of choice, decision time, and confidence. We also need to know how well the model accounts for data in comparison to other possible models that also pass these hurdles. For that reason, we next introduce the Poisson Race model and its account of confidence using Vicker's (1979) balance-of-evidence hypothesis. It is of particular interest because it also provides a different account on the time course of choice and confidence, particularly in the area of subjective probabilities (Merkle & Van Zandt, 2006)

Poisson race model

Diffusion models are a member of a broader class of models called sequential sampling models (Ratcliff & Smith, 2004). A useful distinction between sequential sampling models is based on the stopping rule they employ. Diffusion models use a *relative stopping rule* where the evidence for one alternative in a diffusion model is accumulated in relation to the other alternative. That is, evidence supporting choosing alternative A is evidence against choosing B (Figure 1). As a result, the stopping rule specifies how much evidence supporting one alternative (e.g., A) must exceed the other by a criterion amount – the choice threshold (θ_j) – in order for a judge to make a choice.

An alternative stopping rule is an *absolute stopping rule*. This rule requires evidence in support of each alternative to be stored on separate – often independent – counters and once a counter reaches a criterion level of evidence a choice is made. Thus, evidence for one alternative (e.g., R_A) does not change the amount of evidence in support of the other (e.g., R_B) making the decision process a race between two (or more) counters. The Poisson race model falls into this latter absolute stopping rule class of sequential sampling models and as a consequence gives a different perspective on the time course of choice and confidence (Merkle & Van Zandt, 2006; Van Zandt, 2000b; Van Zandt & Maldonado-Molina, 2004). In what follows we give a brief description of the model. Appendix B lists the response probabilities, the density and cumulative density functions of decision times, and the distribution of confidence ratings for the Poisson race model. The reader is referred to Townsend and Ashby (1983) and/or Van Zandt, Colonius, and Proctor (2000) for a full development of the model as well as Van Zandt (2000b) for the derivation of the confidence distributions.

The basic model assumes that when a judge is presented with a choice between two alternatives, evidence in support of response R_A begins to arrive and increment a counter keeping track of the evidence in support of it. A similar process occurs for response R_B. Rather than using this information to accrue continuous amounts of evidence for each alternative, judges use this evidence to accumulate discrete counts of supporting evidence, much like the firing a of neuron (a discrete event) only occurs after a continuous build up of excitation has reached a criterion level at the synapse (Townsend & Ashby, 1983). The time between the arrival of these counts (inter-arrival times) is continuous and exponentially distributed with the evidence for each alternative arriving at different accrual rates v_i, $i = A, B$. If alternative or stimulus S_A is correct then we expect that its rate should be faster than for the incorrect alternative S_B, $v_A > v_B$. The rates of evidence accumulation serve much the same function as the drift rate in diffusion theory where stimuli that are easy to discriminate between have a greater difference in the rates of evidence accumulation.

As evidence arrives, the two counters race towards a threshold level of evidence. The most general form of this model allows each alternative to have its own threshold level, K_i. These counters operate independently and in parallel. Furthermore, because the inter-arrival times are independent and exponentially distributed, the counts on each counter, $X_A(t)$ and $X_B(t)$, are Poisson processes. As a result, at any given point in time, the probability that the next count

is for alternative A is $v_A/(v_A + v_B)$ and the probability it is for alternative B is $v_B/(v_A + v_B)$ (see Townsend & Ashby, 1983, p. 274). Using these probabilities and the fact that we know the value of the winning counter when a choice is made (e.g., $X_A(t) = K_A$ when alternative A is chosen) we can find the response probabilities $P(R_A|S_A)$ and $P(R_A|S_B)$ or the probability of K_A successes occurring given the respective stimuli. The conditional decision times can also be found as the sum of the exponentially distributed inter-arrival times of evidence until one counter hits threshold (see Appendix B).

A major limitation though of the Poisson race model is that there appears to be a systematic deviation between the shapes of the predicted and observed decision time distributions (Ratcliff & Smith, 2004). The basic problem is that the Poisson counter model produces decision times that are gamma distributed (Townsend & Ashby, 1983). The gamma distribution has the property that the distribution converges toward a normal distribution as the mean decision time increases. However, empirically just the opposite is usually found – that is the distribution of decision times becomes more positively skewed under conditions producing longer mean decision times (Ratcliff & Smith, 2004). This problem does not appear to exist for diffusion models, which naturally produce greater skew with longer mean times.

Balance of evidence hypothesis

An advantage of the Poisson race model is that it can very easily incorporate Vickers' (1979) balance of evidence hypothesis for confidence where confidence in race models is "the difference between the two totals at the moment a decision is reached or sampling terminated" (Vickers, 2001, p. 151). A topic we will return to in the discussion is that in fact one might understand our 2DSD model as a slight modification to the balance of evidence hypothesis. In the 2DSD model, evidence at any given point in time is already relative to the other alternative (thus the term relative stopping rule). But, unlike Vicker's hypothesis which claims confidence only reflects the evidence at the time of choice, in the 2DSD model confidence reflects the evidence collected at choice and evidence collected after choice. To implement the balance of evidence hypothesis within the Poisson race model the internal or covert confidence value C is scaled directly from the difference between the levels of evidence at the time the winning counter reaches its threshold (e.g., $C = X_A(t_D) - X_B(t_D) = K_A - X_B(t_D)$ if A wins) (Van Zandt, 2000b).[8] Realized within the Poisson race model, this hypothesis also means that C is a linear

function of the total amount of evidence accumulated across the two counters and therefore probability distribution of *C* can also be found (see Appendix B).

Vickers and colleagues (Vickers, 1979, 2001; Vickers & Smith, 1985; Vickers, Smith et al., 1985) have shown that with the balance of evidence hypothesis, race models, like the Poisson race model, can account for empirical hurdles 1 to 5 listed in Table 1. The Poisson race model uses the counter thresholds k_j to account for the speed-accuracy tradeoff (hurdle 1) where larger values of the thresholds produce more accurate responses, but longer decision times. In terms of hurdle 2, where confidence is monotonically related to the difficulty of the stimuli, highly discriminable stimuli will have greater differences in the rates of evidence accumulation and thus confidence will increase as discriminability increases. Furthermore, evidence will on average accumulate on both correct and incorrect counters to a greater extent during incorrect trials and thus confidence will be lower on incorrect trials (hurdle 3). In terms of the two-fold relationship between confidence and decision time, during optional stopping tasks, trials with long decision times will be more likely to accrue evidence on both counters and thus will have lower confidence levels than fast trials (hurdle 4). But, by increasing the choice thresholds (either via time pressure or an external force in an interrogation task), the difference between counters will grow and thus confidence will be greater when judges take more time to make a choice (hurdle 5).

A difficulty with the balance of evidence hypothesis implemented in the Poisson race model is the discrete nature of the confidence scale. This limits the model's ability to handle problems related to mapping an overt confidence rating to a covert feeling of confidence. As a result the model has no principled method to account for changes in scales that might occur across conditions (such as can occur when time pressure is manipulated between sessions in combination with the instructions to spread confidence ratings out). This problem also causes trouble for Merkle and Van Zandt's (2006) relative balance of evidence hypothesis, which is of special interest because it has been applied to the subjective probability scale used in the study described later in this article

Relative balance of evidence hypothesis (Merkle & Van Zandt, 2006). According to this hypothesis, after making a choice, a judge's internal subjective belief, *SB*, is given by the expression

$$SB = \frac{K_{win}}{K_{win}+X_{lose}(dt)}. \quad (17)$$

Judges then select a confidence rating $Conf_i$ (.50, …, 1.00) by rounding *SB* to the nearest item on the scale. This method bears some resemblance to ideas in support theory (Brenner, 2003; Rottenstreich & Tversky, 1997; Tversky & Koehler, 1994).

An advantage, in general, of using the Poisson race model with the balance or relative balance of evidence hypothesis to model confidence judgments is that in principle only four or five free parameters are needed to model choice, decision time and confidence distributions. The top four rows of Table 3 lists the free parameters of the model. However, as was previously mentioned, the discrete nature of the internal scale is problematic. Consider, for example, typical values of the choice thresholds $K_A = K_B = 5$ (see for example Ratcliff & Smith, 2004) and the resulting scale produced with the relative balance of evidence hypothesis (Equation 17). These thresholds produce a subjective probability scale where the only possible ratings are 1.00, .80, .70 and .60 (produced respectively by inserting $X_{lose} = 0$, 1, 2, 3, or 4 into the formula for SB, and then rounding to the nearest rating scale value for .50, .60, …, 1.00). In other words, this model predicts that people never select the confidence ratings .50 or .90, which is false. Trial-variability in the choice thresholds (K_{win} and K_{lose}) is one way to work around this problem, as Merkle and Van Zandt (2006) used. Yet, this auxiliary assumption appears to introduce more problems. Consider the case where the number of response categories is varied across conditions, from say a 6 point half-range confidence scale (.50, .60., …, 1.00) to an 11 point full-range confidence scale (.00, .10, …, 1.00). To accommodate the change in scales, the Poisson counter model would need to change the threshold variability in a very ad hoc and psychologically implausible manner.

--

Table 3

One solution to this problem of the finite and discrete confidence scale might be to allow the discrete counts to approach a continuous scale by forcing the model to take a large number of counts before making a choice (e.g., $K_A = K_B = 100$). The problem with this fix is that as noted earlier the decision time distributions are gamma distributed and increasing the number of counts would make the predicted decision time distributions more symmetrical, which is not consistent with empirical data (Ratcliff & Smith, 2004). Consequently, we explored an alternative more

general response mapping hypothesis for the Poisson race model that incorporates a method for mapping the covert balance of evidence in the model to an overt scale using confidence criteria.

Balance of evidence mapping hypothesis. Under this hypothesis we retained the discrete counts, but we posited that the balance of evidence counts, *C*, are perturbed with noise forming a larger population of graded levels of covert confidence, Ω. That is, there is variability in judges' ability to identify different levels of covert confidence. To estimate the distribution over Ω we assumed that there is a Gaussian density centered over *C* so that height of the density is a weighted average of all the observations in the sample. This is akin to using a Gaussian Kernel to estimate a density function in statistics (Van Zandt, 2000a). The height of the estimate is given by

Equation 18 $$f(\Omega) = \frac{1}{h}\sum_c Pr(C = c|R_A, S_A)\phi\left(\frac{c-\Omega}{h}\right)$$

Where $Pr(C = c|R_A, S_A)$is the distribution of confidence levels C under the balance of evidence hypothesis (see Equation B 8**),** ϕ is the standard normal Gaussian density and the free parameter *h* specifies the level of perturbation of C. Figure 3 illustrates this hypothesis for a correct trial in the Poisson model with different levels of the bandwidth parameter *h*. The figure also illustrates that akin to signal detection theory we assume judges partition their covert confidence Ω into the different confidence ratings. Thus, the model handles changes in confidence scales much the same as in signal detection theory and does not necessarily suffer from the same limitations as the other versions of the Poisson race model. Next we evaluate how well these different Poisson race models account for the data in comparison to our 2DSD model.

--

Insert Figure 3

--

Overview of empirical evaluation of 2DSD model

Besides showing that the 2DSD model can account for a wide range of empirical phenomena (hurdles 1 through 5), we also need to address predictions that potentially could invalidate the theory and examine quantitatively how well the model accounts for the data in comparison to competing hypotheses (Roberts & Pashler, 2000). To meet these goals 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. We used the data from

this study to investigate a number of predictions from the 2DSD model that, at first glance, seem unlikely. For example, as Figure 2 and Equations 6 and 7 suggest, the 2DSD model predicts that all else being equal increases in τ should increase both the mean difference between confidence ratings in correct responses and incorrect responses (*slope*) and the variance of the distribution of confidence ratings conditional on being correct and incorrect (*scatter*). In addition, we also examined how well the 2DSD model simultaneously accounts for choice, decision time, and confidence, in comparison to two different versions of the Poisson race model (relative balance of evidence and balance of evidence mapping hypothesis). Finally, we use the 2DSD model to help give a process account of the effect of time pressure on the accuracy of subjective probability estimates in both perceptual and general knowledge tasks.

During the study, six participants completed a large number of trials of the 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 basic task has been studied in a number of studies on confidence and calibration (Baranski & Petrusic, 1994, 1998; Juslin & Olsson, 1997; Vickers & Packer, 1982). During the general knowledge 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, & Kleinboelting, 1991; Juslin, 1994; McClelland & Bolger, 1994). Thus, it is of interest to identify the degree to which the same process can account for judgment and decision processes across these two tasks (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, the 2DSD model 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 desired. In either case, 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. This situation mimics the past empirical studies on the effect of the speed/accuracy tradeoff on confidence used in the past (Baranski & Petrusic, 1998; Vickers & Packer, 1982), which in turn helps assess the degree our model can account for data that was often interpreted in terms of supporting race and/or counter models and the balance of evidence hypothesis. 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 no doubt 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 on a Dell 19 inch flat panel monitor (1908FP) with a NVIDIA GeForce 8600 video card 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 paired comparison experimental task, participants completed a response entry task where they entered a sequence of responses (e.g, 'H', '←', '60'). The task not only helped participants practice the choice and confidence key locations, but also tested the degree to which there was any relationship between the confidence button pressing times (inter-judgment times) and the button locations. During the task, a screen was shown telling people to prepare 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 as quickly as possible the appropriate confidence key. The task was programmed so that participants entered each response sequence (choice and confidence) twice for a total of 24 trials. Note accuracy was not enforced. However, due to a programming error during the line length sessions participants 1 through 4 were forced to enter the correct button. Consequently, these participants completed an extra set of trials of 30 trials per response sequence during an extra session. Table 4 lists the across participants mean time taken to enter the different confidence ratings. There was no clear relationship between inter-judgment times and the confidence level button.

Insert Table 4

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 marker 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 marker with 1 mm space between the central maker and the start of each line. All lines including the central maker were approximately 0.35 mm wide. The central marker and lines were contained in a JPEG high resolution image to insure consistency across monitors. 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. The lengths were chosen to capture a large range of difficulty and confidence ratings.

Design and procedure. For the line length discrimination task, participants completed 10 consecutive experimental sessions. 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. Following Vickers and Packer (1982), within each session participants alternated between blocks of trials where they had a goal of entering an accurate choice and another block of trials where they had a goal of entering a fast choice. In all, they 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). 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. Again across all blocks of trials, 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 of the goal during the current block of trials (speed or accuracy), and (c) at the bottom number 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 (stored as a jpeg image) 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%). Then if they were in the accuracy block feedback was given if they made an incorrect choice and if they were in the speed block feedback was given if they were too slow, otherwise no feedback was given and participants began the next trial.

At the beginning of the experiment, participants were given typical calibration study instructions (cf., Lichtenstein et al., 1982) where they were told to select a confidence rating so that over the long run the proportion 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. As further motivation, across all trials participants earned points based on the accuracy of their choice and confidence rating according to the quadratic scoring rule,

$$points = 100[1 - (correct_i - conf_i)^2] \quad (19)$$

(Stäel von Holstein, 1970). Where $correct_i$ is equal to 1 if the choice on trial i was correct otherwise 0 and $conf_i$ 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. 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 as follows. We obtained a list of the 100 most populated U.S. cities from the 2006 U. S. Census estimates. All possible pairs were formed from these 100 cities producing 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. When the city pairs were shown in the center of a black screen they were centered on

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 for each session 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 speed 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 predictions that the 2DSD model makes regarding the effect of changes in inter-judgment time τ. The second section compares the fit of the 2DSD model to the different Poisson race models (relative balance of evidence hypothesis and balance-of-evidence mapping hypothesis). 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. Finally, we use the 2DSD model to offer a more precise process account of how time pressure can influence the accuracy of subjective probability estimates.

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 decision times were greater than 10 seconds and observed inter-judgment times were greater than 5 seconds. These cutoffs eliminated on average 1.1% (min = 0.03%; max = 6.0%) of the data in the line task and 1.4% (min = 0; max = 5.9 %) of the data in the city task.

Between condition results

Table 5 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).

Insert Table 5

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 between conditions there is a direct relationship between confidence and decision time. 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 θ (Equation 14).

All else, however, did not remain equal. In fact, as Table 5 also shows judges on average increased their inter-judgment time during the speed conditions. In words, judges appear to

compensate for their time pressure when making a choice by taking a little longer to rate their confidence. This is true for both the line and city tasks. Similar results are reported in Baranski and Petrusic (1998). According to the model, this increase in inter-judgment time 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 collecting evidence after making a choice or engage in post decisional stimulus processing to enter their confidence rating (cf., Baranski & Petrusic, 1998). If no post decisional processing occurred and instead choice and confidence simultaneously occurred as most models assume, then one would expect no difference in inter-judgment time between time pressure conditions. In fact, later we show that 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 γ ordinal measure of association (Goodman & Kruskal, 1953, 1954). Goodman and Kruskal's γ only assumes an ordinal scale and makes no distribution assumptions (Goodman & Kruskal, 1953). In addition, unlike Pearson's r, γ 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 (see also Nelson, 1984, 1987).

Table 6 lists the average Goodman and Kruskal's γ between the measures of cognitive performance in each of the tasks across participants. The values below the diagonal are the γ 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 γ coefficient. The associations listed in Table 6 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 6 also shows that confidence is monotonically related to accuracy (hurdle 3) and inversely related to decision time (hurdle 4).

Insert Table 6

Notice also the pattern of γs 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 much 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 6 also reveal places where the 2DSD model is silent. In particular, in all conditions and all tasks inter-judgment time and the confidence rating were negatively related (γ = -.19 to -.54). In other words, when a person gave a higher confidence rating to trial *a* than to *b,* 60 to 77% of the time their inter-judgment time was faster for *a*. This replicates similar results reported by Baranski and Petrusic (1998) and Petrusic and Baranski (2003). The 2DSD model does not explicitly predict such a relationship instead inter-judgment time is an exogenous variable in the model treating the process of selecting a confidence rating much like an interrogation task where judges interrupt the evidence accumulation process. This is a potential weakness of the model and in the discussion we will offer an alternative framing of the 2DSD model treating the confidence process more along the lines of an optional stopping process. This alternative framing can, in turn, offer a process level account of inter-judgment times for each confidence rating.

Finally, the associations in Table 6 are also revealing in terms of the accuracy of subjective probability estimates. Namely in both tasks and in both conditions one of the largest correlations was between accuracy and the confidence rating (γ = .43 to .76). These values indicate that when participants were more confident in their choice for trial *a* than for trial *b* they

were in fact correct 72 to 88% of the time. 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 ($\gamma = .76$ vs. .67, $p < .05$) and city population ($\gamma = .55$ vs. .43, $p < .05$). The 2DSD model attributes this increased resolution of confidence during time pressure to the increase in inter-judgment time. Next we better detail this prediction.

Changes in the distribution of confidence ratings

Everything else being equal the 2DSD model predicts that as inter-judgment time τ increases the mean and the variance of the distribution of evidence used to estimate confidence $L(t_C)$ (see Equation 14 Equation 15) should increase. These changes in the average state of evidence imply that the *slope* score (Yates, 1990),

$$Slope = \overline{conf}_{correct} - \overline{conf}_{error}, \quad (20)$$

should increase as inter-judgment time τ increases. The slope 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 7 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, consistent with the prediction of the 2DSD model regarding the effect of an increase in inter-judgment time, the confidence participants expressed had better unstandardized discrimination between correct and incorrect choices during the speed conditions as opposed to the accuracy conditions.

--

Insert Table 7

--

The increase in slope happened despite the presumably lower choice thresholds in the speed conditions. This is because lower choice thresholds (θ) in the speed conditions lead to a *decrease in both* the average confidence in corrects and incorrects. Recall, however, according to the 2DSD model, increases in inter-judgment time τ lead to an *increase* in the confidence in corrects and a *decrease* in the confidence in incorrects. Thus, the combined effect of lower

choice thresholds ($\downarrow\theta$) and greater inter-judgment times ($\uparrow\tau$) in the speed condition should lead to (a) little change in the confidence in corrects between speed and accuracy and (b) a substantial decrease in the average confidence in incorrects. Indeed empirically this was the case. In the line length task the average confidence in corrects went from .98 (*SE* = .005; Std_{btwn} =.06) in the accuracy condition to .94 (*SE* = .009; Std_{btwn} .07) in the speed condition. In comparison, the average confidence in incorrects went from .81 (*SE* = .003; Std_{btwn} =.11) to .69 (*SE* = .003; Std_{btwn} .09). A similar pattern occurred in the city population task. The average confidence in corrects went from.85 (*SE* = .001; Std_{btwn} =.10) in the accuracy condition to .82 (*SE* = .002; Std_{btwn} .10) in the speed condition. In comparison, the average confidence in incorrects went from .71 (*SE* = .002; Std_{btwn} =.10) to .63 (*SE* = .002; Std_{btwn} .07). 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.

Similar to the slope score, we can also calculate the *scatter* score (Yates, 1990) or the average variance conditional correct and incorrect choices,

$$Scatter = \frac{n_{correct}var(conf_{correct}) - n_{incorrect}var(conf_{incorrect})}{n_{correct} + n_{incorrect}}. \quad (21)$$

According to the 2DSD model 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. Table 7 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) as well as the average participant scatter was significantly larger during the speed condition as compared to the accuracy condition. Only for Partcipant 1 during the city population discrimination task did scatter significantly decrease under time pressure, but note this participant had the lowest levels of proportion correct during this task (60 to 64% correct see Table 2) suggesting this task was fairly difficult for this person. Overall, these results imply that while resolution of confidence ratings increased during the speed conditions so did the variability of the ratings. 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.

While these changes in slope and scatter across speed and accuracy conditions are at least consistent with the 2DSD model, the Poisson race model predicts the opposite pattern: Judges

should be less able to distinguish between corrects and incorrects (slope) with their confidence ratings and are more consistent (scatter) in their ratings under speed as compared to accuracy conditions. That is, the Poisson model predicts the opposite pattern that with lower choice thresholds in speed conditions, the slope and scatter of confidence ratings should be lower. This is because with any Poisson model using a balance of evidence hypothesis (include the relative balance of evidence and the mapping hypothesis) confidence is a direct function of the difference of the amount of evidence accumulated on the two counters (see Appendix B). In the speed condition, the difference between the two counters will be smaller due to the time demands which in turn will reduce the average slope and scatter score. Thus, the slope and scatter differences at least offer preliminary evidence that the Poisson race model using any version of the balance of evidence hypothesis is inconsistent with the data at the individual and group level. A different test of these two models is to assess how well the models account for correct and incorrect decision time and confidence distributions.

Model fitting and comparison

Quantile maximum likelihood method for decision times and confidence. After removing contaminant trials, the raw data in the line length task contains approximately 5,696 total trials with observed choices, decision times, confidence, and inter-judgment times per person. In the city population task this number is 3,826. This high number of trials made it possible to fit the models not just at the level of means but at the level of distributions. In principle, we could fit both the 2DSD model and the Poisson race models directly to the responses 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, Brown, and Mewhort's (2002) quantile maximum likelihood (QML) estimation method to simultaneously fit models to decision time and confidence rating distributions for corrects and incorrects.[9] In the QML 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 QML 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 $L = L_{speed,\ correct} \times L_{speed,\ error} \times L_{accuracy,\ correct} \times L_{accuracy,\ error}$.

If we expand the calculation to also simultaneously incorporate the distribution of confidence ratings, then the data representation is then a 6 (decision time categories) x 6 (confidence ratings) data matrix for corrects and incorrects for both tasks. In principle, one could still use the multinomial distribution function to calculate the likelihood of the 6 x 6 data matrix for corrects and incorrects. This method, however, suffers from a high number of contingencies with 0 cell counts especially in conditions when accuracy is high. This can be problematic for evaluating the fit of models. Instead we calculated the likelihood of the data for the 2DSD model for correct and incorrect responses using the marginal distributions of decision time and confidence ratings.[10] We did the same for the Poisson race model.

Model fitting. In both the line length and city population tasks we broke the data down not only in terms of two levels of time pressure, but also into 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. 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 5), we formed six different levels of difficulty with approximately 300 pairs in each condition.

Thus, the 2DSD and Poisson race 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 QML method, each condition had 10 decision time quantiles + 10 confidence rating frequences = 20 free data points per condition for each of the 10 conditions producing a total of 20×10= 200 data points per participant. The city task with 12 conditions (speed vs. accuracy X 6 levels of difficulty) had 20×12=240 free data points.

In total five model fits are reported for each task. They are summarized in Table 8. The first model was a saturated baseline statistical model. The baseline model used the observed marginal relative frequencies as the predicted probabilities for both the decision time categories and confidence ratings in the QML method. It has an equal number of parameters to free data points therefore it is a saturated statistical model. In terms of the cognitive models, we fit highly constrained models to each participant's data (for a description of the free parameters associated with each model see Table 2 and Table 3). Only the drift rate in the diffusion models and the evidence accrual rate in the Poisson race models were allowed to vary across the levels of difficulty while the choice thresholds were allowed to vary between the time pressure conditions. Although trial-by-trial variability in parameters like the drift rate and the starting point has become a common feature in diffusion models (Ratcliff, 1978; Ratcliff & Rouder, 1998; Ratcliff & Smith, 2004; Ratcliff & Tuerlinckx, 2002; Ratcliff et al., 1999), we felt a comparison between models without incorporating these auxiliary assumptions with trial variability was more informative in these first stages of model evaluation. Later, once we identify the best performing models, we examine the possible improvements trial variability in the process parameters can bring to the models.

--

Insert Table 8

--

Optimization technique. To estimate the maximum likelihood of each of the sequential sampling models in the QML framework we used a variant of Van Zandt's (2000a) iterative Nelder-Mead (Nelder & Mead, 1965) method. During this procedure, the maximum likelihood of the data for a particular model was searched 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 is 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 Townsend & Ashby, 1983; Wagenmakers et al., 2007 for these approximations), and finally another iteration used the best fitting average value across participants from the previous attempts.

Model comparison. Model comparisons were done with the Bayesian Information Criterion (BIC). The BIC is calculated for each model according to the following expression

$$BIC = -2*ML_i + k*\log(N). \quad (22)$$

Where *ML* is the maximum log likelihood of the data for a given model *i, k* is the number of parameters, and *N* is the number of observations (Kass & Raftery, 1995; Raftery, 1995; Wagenmakers, 2007). The model with the smallest BIC is selected as the best fitting model. As a rule of thumb, a BIC difference of 2 or less is interpreted as weak evidence, 2 to 6 as positive evidence, 6 to 10 as strong evidence, and greater than 10 as very strong evidence, for the particular model (Raftery, 1995; Wagenmakers, 2007).

Table 9 and Table 10 list the BIC values for each participant in the line length and city population discrimination task, respectively. The 2DSD and Poisson race models should give a better fit in terms of BIC than the statistical baseline model and do. Across all participants and tasks the worst performing sequential sampling model was the Poisson race model using the relative balance of evidence hypothesis (Equation 17). Its poor performance can be attributed to its inability to simultaneously fit decision times and confidence across the time pressure conditions. For instance, for 4 out of the 6 participants in the speed conditions of the line discrimination task, the maximum likelihood estimates for the choice thresholds were too low (K_{speed} = 4 or lower) to adequately predict the confidence distributions. Another participant, Participant 2, had a value of 6 for the response criteria in the speed and accuracy conditions of the line discrimination task. Thus, the model predicted equivalent performance across the time pressure conditions for this participant in the line discrimination task. This is in direct contrast to the data (see Table 5). A similar result of identical performance predictions occurred for Participants 2, 4, 5, and 6 in the city population tasks.

--

Insert Table 9 and Table 10

--

We interpret this poor performance of the Poisson race model using the relative balance of evidence hypothesis as further evidence of the necessity of any cognitive model of confidence

to explicitly include a method of mapping a covert level of confidence onto an overt scale. The Poisson race model using the balance of evidence mapping hypothesis provides one direct test of this claim. Recall this model assumes the difference between the amounts of information accumulated on the two counters *C* are perturbed by normally distributed noise forming a graded level of covert confidence Ω (see Equation 18). The judge uses confidence criteria to map an overt confidence rating to his or her covert confidence (see Table 8). Fitting this model to the data revealed that across all participants, in both tasks, the Poisson race model using the balance of evidence mapping hypothesis gave a better fit than the relative balance of evidence hypothesis supporting our hypothesis that a method of mapping a covert confidence to an overt response is a necessary component of a complete process model of confidence. The 2DSD model, however, always gave a better fit than either Poisson race model.

To see how and why the 2DSD model provides a better fit we adapted the latency-probability function plots first used by Audley and Pike (1965) and popularized by Vickers (1979) and Ratcliff and colleagues (Ratcliff & Rouder, 1998, 2000; Ratcliff et al., 1999) to include confidence as well. We call the plots latency-confidence-choice probability (LCC) functions. Figure 4 shows the plots for Participant 3 and 4 during the line length task and Figure 5 shows the corresponding plots for the city population task.

--

Insert Figure 4 and Figure 5

--

Overall each LCC function can be understood as a plot showing how the measures of cognitive performance change with stimulus difficulty and are useful for discriminating between sequential sampling models. 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 on the x-axis) and incorrect choices (white circles left of choice probability .5 on the x-axis) 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 (>.9 in the line length and > .8 in the city population task) 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 typically the slowest 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 circles are the data, the dotted lines marked with triangles correspond to predicted functions for the Poisson race model using the balance of evidence mapping hypothesis, and the dashed lines marked with squares correspond to the predicted functions for the 2DSD 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 score calculations the average confidence rating in corrects is greater than errors, and this difference between confidence in corrects and incorrects increases as the stimuli get easier (moving out from the confidence-choice probability functions in the upper panels). This is true for both the line length (Figure 4) and city population (Figure 5) discrimination tasks. The 2DSD model gives a good account of the confidence and change in confidence from easy corrects to easy incorrect. In comparison, the Poisson race model using the balance of evidence mapping hypothesis only gives a good account of confidence during the accuracy conditions, but under predicts the slope of the confidence ratings during speed conditions. This is because, as explained earlier, the Poisson race model will on average predict lower slopes for the speed condition. Across participants we found the opposite result. The slope increases from accuracy to speed (see also Table 7). The 2DSD model attributes this result to participants' increased inter-judgment time under time pressure.

The LCC functions also expose the ability of the sequential sampling models to account for (or to fail to account for as it may be) decision times. For example, a well known property of diffusion models like the 2DSD model is the symmetrical prediction of decision times for corrects and incorrects evident with the inverted U shape of the latency-choice probability functions in the lower half of the panels (see Figure 4 and Figure 5) (Ratcliff & Rouder, 1998; Ratcliff et al., 1999). As is also evident, accumulator models like the Poisson race model with an absolute criterion predict a monotonically increasing function from fast easy corrects to slow easy incorrects. In both the line length (Figure 4) and city population (Figure 5) tasks the decision time data appear to conform closer to the best fitting functions of the 2DSD model though there are certainly systematic departures. For example, they show that slow errors

occurred where decision times were slower for errors than correct responses for the corresponding level of difficulty. Consistent with past findings (Estes & Wessel, 1966; Luce, 1986; Ratcliff & Rouder, 1998; Swensson, 1972; Townsend & Ashby, 1983), the slow errors are especially evident when accuracy was emphasized and during the more difficult conditions. Consistent with this fact, slow errors become even more evident during the more difficult city population task (see Figure 5). We have added the result of slow errors as the sixth empirical hurdle in Table 1 that any complete model of cognitive performance must account for.

Interestingly, 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). While this pattern is not strongly present in our data, we have also added this result as hurdle 7 to Table 1. Ratcliff and colleagues have shown that diffusion models which assume trial-by-trial variability in the drift rate and in the starting point z can simultaneously account for slow errors (hurdle 6) and fast errors (hurdle 7) (Ratcliff & Rouder, 1998; Ratcliff et al., 1999). In the next section we briefly explore the consequences of this addition in the 2DSD model. Given the inability of the Poisson race model to qualitatively and quantitatively account for the data in a parsimonious manner, we did not pursue further the consequences of trial variability in the Poisson race model.

Trial-by-trial Variability in Process Parameters

A widely accepted assumption of stochastic models is that from trial to trial stimuli – even nominally identical stimuli – are not always processed identically. This is true for Thurstonian scaling (Thurstone, 1927), signal detection theory (Green & Swets, 1966), and for diffusion models (Ratcliff, 1978; Ratcliff & Rouder, 1998; Ratcliff & Smith, 2004; Ratcliff et al., 1999). Trial-by-trial or trial variability within diffusion models makes clear and identifiable predictions. Intuitively, one source of trial variability, perhaps due to attentional lapses, is in the quality of the evidence accumulated or the drift rate δ. To model this we assume the drift rate for stimuli in the same experimental conditions is normally distributed with a mean v and a standard deviation η. Variability in the drift rate across trials causes the diffusion model to predict slower decision times for incorrect choices than for correct choices (Ratcliff & Rouder, 1998) (hurdle 6).

We also incorporated a second type of variability in the form of the starting point (z) variability where the start point varies across trials in the form of a rectangular distribution with mean z and range s_z. Variability in the starting point across trials causes the diffusion model to predict faster decision times for incorrect choice than for correct choices, especially when speed is emphasized (Ratcliff & Rouder, 1998; Ratcliff et al., 1999). Thus, in total, these two additional auxiliary assumptions add two additional process parameters to the 2DSD model, but allow for the same diffusion process to account simultaneously for slow and fast errors (Ratcliff & Rouder, 1998).

To examine whether these auxiliary assumptions of trial variability aided in our ability to account for the data, we fit the 2DSD model with trial variability in the drift rate and starting point to the data from each task for each participant using the adapted QLM method. A similar procedure was employed where the mean drift rate was allowed to vary across the 5 or 6 different levels of difficulty and only the choice threshold was allowed to vary across the different time pressure conditions. Table 11 lists the BIC values for the 2DSD model with drift rate and starting point variability as well as the previous BIC values from the 2DSD model without trial variability in the parameters. The model fits reveal that for everyone save Participant 2 in the city population task, trial variability improved the fit of the 2DSD mode (see Appendix C for parameter estimates). Figure 6 and Figure 7 display the fit of the 2DSD model with trial variability in terms of LCC functions. They show, in particular, the role trial variability has in improving the ability of the models to account for incorrect decision times that are on average slower than the decision times for the correct choices in the same condition (slow errors). The LCC functions also reveal that, at least at the level of means, trial variability does not appear to aid in the 2DSD model's account of confidence ratings.

Insert Table 11, Figure 6, Figure 7

Trial variability does, however, account for the observed relationship between confidence and decision times even when stimulus difficulty is held constant. The 2DSD model (without trial variability) predicts that for a given stimulus decision times and confidence are independent of each other for correct and incorrect choices. 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 12 lists the average Goodman and Kruskal γ rank order correlation between decision time and confidence for each participant in the line length task holding the stimulus difficulty constant. Note Participant 5 was excluded from this calculation as he or she used very few confidence ratings in this task (as his or her best fitting criterion estimates in Appendix C reveal). 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.

--

Insert Table 12

The 2DSD model with trial variability attributes this relationship to trial by trial variability in the drift rate. That is, even when stimulus difficulty is held constant there is an inverse relationship between decision time and confidence (hurdle 3 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. To illustrate this effect, we calculated Participant 2's best fitting joint density function of observed decision time $t_D' = t_D + t_{ED}$ and the accumulated evidence at the time of confidence $L(t_C)$ for correct responses during the accuracy conditions marginalized across the possible drift rates for the fourth level of difficulty (32 vs 33.87 mm). Figure 8 shows the contour plot of this joint distribution. The horizontal lines across the figure indicate the location of the criteria for the different confidence ratings. The contour plot of the joint distribution shows that with trial variability in the drift rate, the 2DSD model can account for the inverse relationship between decision time and confidence even when difficulty is held constant. That is, the reason for the inverse relationship between decision time and confidence when difficulty is held constant is because of trial variability in the quality of the processing of stimuli perhaps due to attentional lapses. The model does, however, tend to underestimate the relationship between observed decision time and confidence. For example, the observed Goodman and Kruskal γ between decision time and confidence for Participant 2 in this level of difficulty was -.32 (SE = .04) while the predicted correlation was γ = -.24. One explanation for

the underestimation is that the models, as explained in the model fitting section, are fit to the marginal distributions of confidence and decision times.

--

Insert Figure 8

--

Accuracy of confidence ratings

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 (de Finetti, 1937; Savage, 1954). To that end, the 2DSD model can be used to understand how people make subjective probability estimates and the time course of that process. There are, however, two different dimensions by which a subjective probability estimate can be evaluated: substantively and normatively (Winkler & Murphy, 1968).

Substantive goodness

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 earlier, slope scores increased under time pressure in both perceptual and general knowledge tasks (Table 7) indicating an enhanced resolution about the accuracy of their choice under time pressure experienced during choice. The 2DSD model attributes this enhanced resolution to the increase in inter-judgment time judges allocate to form a confidence rating when experiencing time pressure at choice.

Recall, that there is also a difference in scatter or the averaged variance conditional on correct and incorrect choices, which is again consistent with the 2DSD model (Table 7). The increase in scatter, however, may detract from the increase in slope in terms of a judge's resolution (Wallsten, Budescu, Erev, & Diederich, 1997; Yates & Curley, 1985). To examine this question we calculated $DI' = slope / \sqrt{scatter}$ so that the mean difference between confidence ratings for corrects and incorrects is taken relative to a common standard deviation (Wallsten et al., 1997). Table 13 lists the mean and standard deviations of the DI' scores across participants in each task (individual estimates can calculated using the values in Table 7). 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 is in fact consistent with the 2DSD model because in diffusion models in general standardized accuracy grows as a linear function of the square root of time (e.g., $\sqrt{\tau}$) (see Equation 8). Finally, recall that this increase in resolution in the speed conditions is evident even with Goodman and Kruskal's ordinal measure of association γ (Table 6) (see Nelson, 1984, 1987 for an argument as to the use of γ as a primary measure of accuracy). 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).

--

Insert Table 13

--

Normative goodness

Substantive goodness is but one dimension we demand from our forecasters. When confidence ratings come in the form of subjective probabilities, we also demand them to adhere to the properties of probabilities. One reason why is that 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 of personal and notso personal 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 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 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 who uses the long run historical relative frequency of rain during a particular month as 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,

$$bias = \overline{conf} - pc, \tag{23}$$

where $bias > 0$ indicates overconfidence.[11] That is judges tend to overestimate the likelihood they are correct. Table 13 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 5). 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 13 also shows that there was very little statistical difference between the bias scores of the speed and accuracy conditions. This suggests that time pressure during choice had little effect on people's overall calibration, replicating results from Baranski and Petrusic (1994). Although one might expect that miscalibration might increase with time pressure, the 2DSD model can be used to understand the lack of an effect. Note first the model does an adequate job in accounting for the changes in (or lack thereof) of the bias scores. In the line length task, the average proportion of variance accounted for across the 2 time pressure conditions and the 6 levels of difficulty was $R^2 = .85$ with a mean absolute deviation (MAD) between the observed

and predicted bias score of .03. In the city population task, the average proportion of variance account for was R^2 = .74 with a MAD of .03.

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

$$bias = pc(slope - 1) + \overline{conf}_{incorrect}. \quad (25)$$

Where *slope* = $\overline{conf}_{corr} - \overline{conf}_{incorr}$. In this study with the two-choice half scale, the slope score is bounded to be less than .5 ($slope \leq .5$), while the mean confidence in incorrects has to be greater than .5, $\overline{conf}_{incorrect} \geq .5$. Therefore, using Equation 25, judges will move towards more overconfidence (bias > 0) if there is (a) a decrease in the proportion correct, (b) an increase *slope*, and/or (c) increases in $\overline{conf}_{incorrect}$. Recall consistent with the 2DSD model, during the speed condition due to a decrease in θ there was a decrease in the proportion correct (*pc*). There was also an increase in slope (due to an increase in τ). But, also consistent with the model the change in slope was due to an interaction between the decrease in θ and increase in τ so that $\overline{conf}_{incorrect}$ was substantially lower in the speed conditions (whereas $\overline{conf}_{correct}$ was only slightly lower). Taken together the 2DSD model implies that the post-decision processing of evidence and its dependence on τ helps makes the judgment system fairly robust to the influence of time pressure on calibration.

Reliability diagrams or calibration curves are a useful way to see how well the 2DSD model accounts for this correspondence between stated confidence ratings and the proportion of correct inferences. Calibration curves plot the observed or predicted 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 9 shows the calibration plot for the two different tasks averaged across the six participants. The hard conditions represent the hardest three conditions of each task and the easy condition represents the easiest three. The plotted lines represent the 2DSD model with trial variability predictions for the calibration curve for each participant. The calibration curves for the model were calculated using the parameters found to best fit the distributions of the cognitive performance indices. The calibration curves can be found with the following expression for the proportion of correct choices conditional on each confidence rating $conf_j$,

$$P(correct|conf_j) = \frac{P(conf_j|correct)P(correct)}{P(conf_j|correct)P(correct)+P(conf_j|incorrect)P(incorrect)}. \quad (24)$$

Where $p(correct)$ and $p(error)$ are found with Equation 9 and $p(conf_j|correct)$ and $p(conf_j|error)$ with Equation 16. The figures show that the 2DSD model does a pretty good job capturing the changes in calibration across the various conditions. The model does appear to depart from the data slightly for the confidence ratings of .5 and .6 in the accuracy conditions of the line length task. As the error bars indicate, this is likely due to a lower number of observations at these levels of confidence in this condition. For instance, in these conditions, there were on average about 105 observations per participant at the .5 level as compared to 750 at the 1.00 level.

--

Figure 9

--

As is evident in the calibration curves in Figure 9 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; 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 (Std_{btw} = .09; $\overline{SE}$ = .004) in the hardest three conditions to .00 (Std_{btw} = .05; $\overline{SE}$ = .002) in the easiest conditions. In the city population task, these numbers were .15 (Std_{btw} = .07; $\overline{SE}$ = .005) and .02 (Std_{btw} = .06; $\overline{SE}$ = .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 (υ in the model with trial variability and δ 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 & McGoey, 1980; Suantak, Bolger, & Ferrell, 1996).[12]

Overall accuracy

A question that permeates all of these analyses is why did participants respond to changes in time pressure by increasing their inter-judgment time? Intuitively, the reaction of participants seems sensible. When forced to make a choice under time pressure, it seems reasonable to take a bit longer to assess one's confidence in that choice. We also argue that in this two-choice, half scale task this reaction is at least consistent with an optimal solution where a judge wants to minimize decision and inter-judgment time as well as maximize choice and confidence accuracy. Choice and confidence accuracy (or inaccuracy) can be measured with the Brier (1950) score

$$Brier = (conf_i - correct_i)^2 \quad (26)$$

In this equation, $correct_i$ is equal to 1 if the choice on trial i was correct otherwise 0 and $conf_i$ was the confidence rating entered in terms of probability of correct (.5, .6, …, 1.0). 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 & Pfanzagal, 1966; Murphy, 1973; von Winterfeldt & Edwards, 1986; Yates, 1990). This means that if participants have an explicit goal of minimizing their Brier score they will only minimize their Brier score 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 ($conf$). Recall participants were rewarded in both tasks according to a linear transformation of the Brier score (see Equation 19). 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 it can be decomposed in the following manner:

$$\overline{Brier} = VI + Bias^2 + (VI)(Slope)(Slope - 2) + Scatter \quad \textbf{(27)}$$

(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 $VI = p(correct)p(incorrect)$. Thus, according to this decomposition, the Brier score integrates both normative and substantive goodness (cf., Winkler & Murphy, 1968). Furthermore, the decomposition also exposes how changes in the 2DSD model influence the Brier score.

Table 13 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 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 .206 (SE = .002; Std_{Btwn} = .021) in the hardest three conditions which was significantly larger than the average Brier score in the easiest three conditions .047 (SE = .001; Std_{Btwn} = .006). In the city population task, the average Brier score was .253 (SE = .002; Std_{Btwn} = .019) in the hardest three conditions which was significantly larger than the average Brier score in the easiest three conditions .159 (SE = .002; Std_{Btwn} = .033).

The 2DSD model with trial variability gives a good account of the average Brier score across the conditions. In the line length task the average proportion of variance explained (across participants) was R^2 = .98 with an average mean absolute deviation between observed and

predicted Brier score of MAD = .01. In the city population task the proportion of variance explained was R^2 = .85 with an average MAD = .02. Using the Brier score decomposition in Equation 27 we can see how the 2DSD model accounts for the Brier score and changes in the Brier score. For example, increases in slope will by and large decrease the Brier score. This in turn implies that when judges increased their inter-judgment time τ under time pressure (thus lower choice thresholds θ) they were at least in principle counteracting the negative impact of lower choice accuracy on the Brier score and their final earnings.

Taken together these results imply that participants can use their choice thresholds (θ) and the inter-judgment times (τ) to improve their Brier score. As an example, Figure 10 shows the predicted Brier score surface for a range of choice thresholds (θ) and the inter-judgment times (τ) for participant 6 within the 4th level of difficulty in the city population task. For simplicity we used the 2DSD model *without* trial variability to calculate the surface. It shows that the predicted Brier score tracks these two parameters τ and θ and that an increase in either parameter will decrease their Brier score, but increasing either of these parameters will also increase the time taken to make a choice and/or confidence rating. This pattern of trading off accuracy with speed suggests that if judges have a goal to minimize choice and inter-judgment time and minimize the Brier score, then the 2DSD model can give – for a given level of difficulty – the optimal choice threshold θ and inter-judgment time τ to meet this goal. Formally, the 2DSD model specifies a choice threshold θ and inter-judgment time τ that minimizes the cost function

$$Cost = c_1(Decision\ Time + InterJudgment\ Time) + c_2 Brier. \qquad (28)$$

Where c_1 and c_2 specify the cost of time and the level of a Brier score, respectively.

--

Insert Figure 10

--

Returning to the question as to why participants responded to time pressure at choice by increasing their inter-judgment time. The answer at least according to the 2DSD model is that this strategy is consistent with the optimal solution in this task. That is, given that participants sought to minimize choice and inter-judgment time and also minimize the Brier score, then under time pressure at choice, the optimal solution is to increase their inter-judgment time.

Discussion

We have recast the standard diffusion model to give a single process account of choice, decision time, and confidence. Perhaps the model is best summarized by answering a recent question from a colleague curious whether choice, decision time, or confidence, was a purer measure of the judgment process during a two-alternative forced choice task. The 2DSD model 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 information accumulation process.

In terms of the 2DSD model, 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 an addendum to this process where evidence continues to accumulate after a choice. Judges then interrupt the process to categorize the accumulated evidence into 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 a confidence rating is mapped to the total state of evidence at confidence time.

In the discussion that follows we summarize the implications of the 2DSD model for our understanding of the cognitive underpinnings of confidence judgments and our understanding of the accuracy of subjective probability estimates. We conclude with an argument that these two issues, process and accuracy, go hand-in-hand. That is, we cannot and will not understand what makes subjective probability estimates accurate without understanding the processes underlying them. But, at the same time, we cannot and will not understand the process without understanding the goals judges have when making these estimates.

Cognitive underpinnings of confidence

Reframing of past hypotheses

Cognitively, the model reframes two previous hypotheses about confidence. The first reframed hypothesis is Vicker's (1979) balance of evidence hypothesis. In the 2DSD model, confidence is – as Vicker's (1979) originally postulated –a function of the balance of evidence in favor of one alternative over the other. This is a natural product of a diffusion 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 like the

Poisson model, evidence on one counter must be compared to the other counter. Indeed as Vickers and Packer suggested (1982) this difference between counters is the natural analog to evidence in diffusion models. The divergence between race and diffusion models, though, is that the difference needs to be calculated after the decision in race models.

This post-decisional computation led Baranski and Petrusic (1998; see also Petrusic & Baranski, 2003) in part to postulate that there is some post decisional processing of confidence and to show that there is some time that passes after a choice but before a confidence judgment is made that cannot be attributed solely to motor time. Our 2DSD model suggests this post decisional processing is not some sort of computation, but a 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 new phenomena like an increase in the resolution of confidence judgments when participants face time pressure at choice (see Table 1).

An interaction between choice and confidence stages

More importantly, though, the 2DSD model 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 the 2DSD model 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) also cannot easily handle this pattern. The Poisson race model, however, may be able to account for the interaction between choice and confidence stage if it is 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). This two-stage Poisson race model, however, would still have difficulty accounting for differences in confidence scales within and between studies. A solution to this problem is the balance of evidence mapping hypothesis we examined. We leave an assessment of this solution to future research, but do note that this fix would substantially increase the number of parameters in the Poisson race model. In our study, it would require the Poisson race model (without trial variability) to have 17 to 18 parameters.

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 changes depending on the time interval between choice and ratings. 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.

Nevertheless Ratcliff and Starns' (2009) RTCON model does expose a weakness in the 2DSD model: the 2DSD model 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 confidence ratings and the criteria are fixed to specific evidence states regardless of the implicit unobserved 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 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). Using this assumption the model then makes precise and testable predictions regarding the distribution of observed response times and confidence ratings.

Explaining Inter-judgment times

Another limitation of the 2DSD model is that inter-judgment times are an exogenous parameter in the model. However, the correlations in Table 6 reveal that for both the line length and city population tasks there was a negative relationship between (a) the observed inter-judgment time and difficulty; and (b) the observed inter-judgment time and the confidence rating given. Figure 11 displays the average inter-judgment time for the line-length task for the speed

and accuracy conditions. This generally decreasing monotonic relationship between confidence and inter-judgment time has been interpreted as further evidence of some post-decisional computational processing (Baranski & Petrusic, 1998; Petrusic & Baranski, 2003; Van Zandt & Maldonado-Molina, 2004). The 2DSD model, as currently formalized, is silent on this matter.

--

Insert Figure 11

--

With a different stopping rule, however, the 2DSD model can account for these properties of the inter-judgment time. To do so notice that up to this point, the model has used an interrogation-type stopping rule to determine when judges stop accumulating evidence and map a confidence rating to the evidence state. That is some cue, force, or time limit, external to the system determines when judges stop collecting evidence. Alternatively, the second stage can be reformulated as using an optional stopping rule to form a confidence rating, where some standard internal to the 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 the 2DSD model as a Markov chain (e.g., Diederich & Busemeyer, 2003) as shown in Figure 12. The top chain describes the choice process. The circles represent different evidence states ranging from the lower choice threshold $-\theta$ to the upper threshold θ. Evidence accumulation adjusts the judge's evidence state up a step ($+\Delta$) with probability p or down a step ($-\Delta$) with probability q. The evidence states corresponding to the choice thresholds (black circles) denote the typical absorbing barriers in diffusion models where 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 D; Diederich & Busemeyer, 2003).

--

Insert Figure 12

--

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 15 for the chain when Alternative A is

chosen). Now, however, we take a different tract with the 2DSD model. Instead of a fixed time interval assumption, we assume markers are placed along the evidence state space representing the different confidence ratings (.50, .60, ..., .90,1.00), one for each rating. For the confidence ratings below certainty (< 1.00), each time the judge passes one of these markers there is a probability w_{conf} that the judge exits and gives the corresponding confidence rating.[13] Two boundary assumptions for this process were also made. First, the evidence state representing the confidence rating of 1.00 (or certainty) was set equal to an absorbing boundary ($w_{1.00}$ =1.0, thus the black circle shown in the lower chain in Figure 15). This means that once the process enters this state representing confidence level 1.00 or 'certainty', with probability 1 the evidence accumulation process ends and the corresponding confidence rating is given. This assumption was put in place after preliminary fits of the model revealed it necessary to account for the inter-judgment times. Interestingly, this idea of treating the 'certain' response as different from other responses is consistent with some theories of probability like Keynes' (1948) logical relational theory of probability where the estimates of 'certainty' and 'impossibility' hold a special privileged position among all other subjective probabilities such that all other subjective probability estimates (verbal or numeric) fall between these ratings (see p. 38). A second boundary assumption was that we placed a reflecting boundary (black bar in Figure 12) at a sufficient distance below .5 so that the evidence accumulation process was reflected back in the opposite direction much like a ball bouncing off a wall (see Cox & Miller, 1965, p. 24). 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 D).

To explore this model we fit the model using least squares methods to the average participant's choice proportions, mean decision times for corrects and incorrects, relative frequency of confidence ratings for correct and incorrects, and mean inter-judgment times for correct and incorrect choices (see Appendix D). In fitting the model to the tasks, only the choice threshold parameter (θ) and the probabilities of exiting for each confidence rating (w_{conf}) was allowed to vary between the speed and accuracy conditions. In the end, for each task there were 50 data points and 17 free parameters. The model does a reasonably good job of accounting for the data across these four different sets of cognitive performance indices in both tasks. The proportion of variance accounted for in the line length task was R^2 = .72 and in the city population task R^2 = .75.

Focusing on the inter-judgment times, Figure 11 shows the 2DSD confidence marker model can account for both the increased inter-judgment times for the speed condition and the decreasing inter-judgment times associated with greater levels of confidence in the line length task. A similar pattern was evident for the city task though preliminary simulations suggest that trial variability in the drift rate may prove necessary to fully account for the slower inter-judgment times found in this task (Table 5). In terms of the increased inter-judgment times during the speed condition, the model fits indicate that the reason judges showed this pattern was that they lowered their confidence marker probabilities ($w_{.50, .60, \ldots, .90}$) when they faced time pressure at choice and as a result collected more evidence before making a confidence judgment. The average *w* was .17 in the accuracy condition of the line-length task and .02 in the speed condition with similar values in the city population task.

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, by and large, a dissociation was found 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 (Table 13) and their bias decreased as the stimuli got easier. Empirically, 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 (.87 in the speed condition to .95 in the accuracy conditions) than participants in other studies (e.g., .65 to

.68 in study 1 in Keren (1988). But, more importantly, we showed that the 2DSD model can give a reasonably good account of the distributions of cognitive performance indices ranging from choice proportions to decision times to confidence ratings to even inter-judgment times in both tasks. This implies that a single choice and judgment process may underlie both tasks.

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, 2000, 2001, 2007) and these neural mechanisms are embedded in the sensory-motor circuitry (Heekeren et al., 2008; Shadlen & Newsome, 2001; Tosoni, Galati, Romani, & Corbetta, 2008). 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 (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. Our 2DSD model, however, goes beyond these results suggesting that the same mechanism(s) may be used to make confidence judgments and furthermore in tasks where judges first make a choice, there is post-decisional processing of the evidence.

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. This problem of the accuracy of subjective probabilities is an age-old problem in the cognitive and decision sciences. In fact, Pierce and Jastrow in 1884 noted two observations: (a) the level of confidence judges had in a choice tracked their choice discrimination and (b) that on those trials when judges reported guessing they actually did much better than chance. These observations align very well with Winkler and Murphy's (1968) dimensions of substantive (resolution) and normative (calibration) goodness. There have been several descriptive theories as to why, when and how these judgments are accurate or inaccurate ranging from heuristic-based (Tversky & Kahneman, 1974) to memory-based (Dougherty, 2001; Koriat, Lichtenstein, & Fischhoff, 1980; Sieck, Merkle, & Van Zandt, 2007) to environmental (Gigerenzer et al., 1991; Juslin, 1994) to stochastic (Budescu, Erev, & Wallsten, 1997; Erev et al., 1994) to statistical

(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 speak to the effects of time pressure on the accuracy of subjective probabilities.

The 2DSD model, in fact, shows that the time course of confidence judgments can have pervasive effects on all the dimensions of accuracy from the substantive goodness of confidence judgments to the normative goodness of these same judgments to the overall accuracy of the choice and judgments. The 2DSD model largely isolates these effects to changes in inter-judgment time. In particular, when faced with time pressure, increases in inter-judgment time can help judges maintain their normative goodness or the correspondence between subjective probability estimates and the actual relative frequency of events (bias). At the same time, the increase in inter-judgment time also improved the substantive goodness of subjective probabilities revealing that judges can potentially have greater resolution when under time pressure at choice.

Although judges may often face a time pressure situation like the one in this study where there is pressure when making a choice and little pressure when assessing confidence, the 2DSD model also reveals a larger set of time pressure situations than were studied here. One way to conceptualize the larger set is in a factorial design where various levels of time pressure during choice are crossed with various levels of time pressure at confidence assessment. The 2DSD can, in turn, reveal how these different time pressure situations influence the accuracy of subjective probability estimates.

While the accuracy of subjective probability estimates has vast practical implications, there has been a call for basic judgment research to orient away from questions of response accuracy and instead focus more on response distributions (Erev et al., 1994; Wallsten, 1996). The concern is in reaction to studies focusing solely on analyses of the accuracy of confidence judgments in order to uncover the psychological processes underlying confidence judgments. The problem is that a judge's accuracy is not entirely under the judge's control. Changes in stimuli can also change the accuracy of subjective probabilities (see Equation 24). As a result,

changes in accuracy may not necessarily be indicative of changes in psychological processes (Wallsten, 1996). We echo this call, but we also expand on this idea by pointing out that many times accuracy and process go hand-in-hand. For example, our analysis with the 2DSD model suggests that accuracy has a direct impact on the judgment process. That is, if the goal is to minimize choice and inter-judgment time and maximize choice and confidence accuracy (see Equation 28), then under time pressure at choice the optimal solution is to increase inter-judgment time. In other words, without understanding the role accuracy plays in behavior 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 the 2DSD model. Thus, accuracy and process must be understood in tandem. If one wants to understand one then the other must be understood as well.

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). The 2DSD model 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 the 2DSD model 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? Or what are the neural mechanisms underlying confidence judgments, are they the same as those underlying decision? We think the 2DSD model provides a useful framework for taking on these larger and more difficult questions.

References

Aczel, J., & Pfanzagal, J. (1966). Remarks on the measurement of subjective probability and information. *Metrika, 11*, 91-105.

Adams, J. K. (1957). A confidence scale defined in terms of expected percentages. *The American Journal of Psychology, 70*(3), 432-436.

Adams, J. K., & Adams, P. A. (1961). Realism of confidence judgments. *Psychological Review, 68*(1), 33-45.

Ariely, D., Tung Au, W., Bender, R. H., Budescu, D. V., Dietz, C. B., Gu, H., et al. (2000). The effects of averaging subjective probability estimates between and within judges. *Journal of Experimental Psychology: Applied, 6*(2), 130-147.

Ascher, D. (1974). *A model for confidence judgments in choice tasks.* McMaster University, Hamilton, Ontario.

Ashby, F. G. (1983). A Biased Random-Walk Model for 2 Choice Reaction-Times. *Journal of Mathematical Psychology, 27*(3), 277-297.

Ashby, F. G. (2000). A stochastic version of general recognition theory. *Journal of Mathematical Psychology, 44*(2), 310-329.

Audley, R. J. (1960). A stochastic model for individual choice behavior. *Psychological Review, 67*, 1-15.

Audley, R. J., & Pike, A. R. (1965). Some alternative stochastic models of choice. *British Journal of Mathematical & Statistical Psychology, 18*(2), 207-225.

Baranski, J. V., & Petrusic, W. M. (1994). The calibration and resolution of confidence in perceptual judgments. *Perception & Psychophysics, 55*(4), 412-428.

Baranski, J. V., & Petrusic, W. M. (1995). On the Calibration of Knowledge and Perception. *Canadian Journal of Experimental Psychology-Revue Canadienne De Psychologie Experimentale, 49*(3), 397-407.

Baranski, J. V., & Petrusic, W. M. (1998). Probing the locus of confidence judgments: Experiments on the time to determine confidence. *Journal of Experimental Psychology: Human Perception and Performance, 24*(3), 929-945.

Bjorkman, M., Juslin, P., & Winman, A. (1993). Realism of Confidence in Sensory Discrimination - the Underconfidence Phenomenon. *Perception & Psychophysics, 54*(1), 75-81.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review, 113*(4), 700-765.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78, 1-3.*

Brenner, L. A. (2003). A random support model of the calibration of subjective probabilities. *Organizational Behavior & Human Decision Processes, 90*(1), 87-110.

Budescu, D. V., Erev, I., & Wallsten, T. S. (1997). On the importance of random error in the study of probability judgment: Part I. New theoretical developments. *Journal of Behavioral Decision Making, 10*(3), 157-171.

Budescu, D. V., Erev, I., Wallsten, T. S., & Yates, J. F. (1997). Introduction to this special issue on stochastic and cognitive models of confidence. *Journal of Behavioral Decision Making, 10*(3), 153-155.

Budescu, D. V., & Wallsten, T. S. (1995). Processing linguisitic probabilities: Geneeral principles and empirical evidence. In Busemeyer, R. Hestie & D. Medin (Eds.), *Decision*

Making from the Perspetive of Cognitive Psychology (pp. 275-318). New York: Academic press.

Busemeyer, J. R., & Diederich, A. (2009). *Methods for Cognitive Modeling*. New York: Sage Publications

Busemeyer, J. R., & Goldstein, D. (1992). Linking together different measures of preference: A dynamic model of matching derived from decision field theory. *Organizational Behavior & Human Decision Processes, 52*, 370-396.

Busemeyer, J. R., & Townsend, J. T. (1992). Fundamental derivations from decision field theory. *Mathematical Social Sciences, 23*, 255-282.

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review, 100*(3), 432-459.

Cox, D. R., & Miller, H. D. (1965). *The Theory of Stochastic Processes*. New York, NY: Chapman and Hall.

Dawes, R. M. (1980). Confidence in intellecutal judgments vs. confidence in perceptual judgments. In E. D. Lantermann & H. Feger (Eds.), *Similarity and Choice: Papers in honour of Clyde Coombs*. Bern, Switzerland: Hans Humber Publishers.

de Finetti, B. (1937). La prevision: Ses lois logiques, sets sources subjectives. *Annales de l'Institut Henri Poincaré*.

Diederich, A. (1997). Dynamic Stochastic Models for Decision Making under Time Constraints. *Journal of Mathematical Psychology, 41*(3), 260.

Diederich, A., & Busemeyer, J. R. (2003). Simple matrix methods for analyzing diffusion models of choice probability, choice response time, and simple response time. *Journal of Mathematical Psychology, 47*(3), 304-322.

Dougherty, M. R. P. (2001). Integration of the ecological and error models of overconfidence using a multiple-trace memory model. *Journal of Experimental Psychology: General, 130*(4), 579-599.

Edwards, W. (1965). Optimal Strategies for Seeking Information - Models for Statistics, Choice Reaction-Times, and Human Information-Processing. *Journal of Mathematical Psychology, 2*(2), 312-329.

Egan, J. P. (1958). *Recognition memory and the operating characteristic (Tech. Rep. No. AFCRC-TN-58-51)*. Bloomington, IN: Indiana University, Hearing and Communication Laboratoryo. Document Number)

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review, 101*(3), 519-527.

Estes, W.K., & Wessel, D. L. (1966) Reaction time in relation to display size and correctness of response in forced-choice visual signal detection. *Perception and Psychophysics*, 1, 369-373.

Feller, W. (1968). *An introduction to probability theory and its applications*. New York: Wiley.

Ferrell, W. R., & McGoey, P. J. (1980). A model of calibration for subjective probabilities. *Organizational Behavior & Human Decision Processes, 26*(1), 32-53.

Festinger, L. (1943). Studies in decision: I. Decision-time, relative frequency of judgment, and subjective confidence as related to physical stimulus difference. *Journal of Experimental Psychology, 32*, 291-306.

Feynman, R. P. (1986). *Appendix F of Rogers Commission Report: Personal observations on the reliability of the shuttle*. Retrieved. from.

Garrett, H. E. (1922). A study of the relation of accuracy to speed. *Archives of Psychology, 56*, 1-105.

Gigerenzer, G., Hoffrage, U., & Kleinboelting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98*(4), 506-528.

Gold, J. I., & Shadlen, M. N. (2000). Representation of a perceptual decision in developing oculomotor commands. *Nature, 404*(6776), 390-394.

Gold, J. I., & Shadlen, M. N. (2001). Neural computations that underlie decisions about sensory stimuli. *Trends in Cognitive Sciences, 5*(1), 10-16.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience, 30*, 535-574.

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review, 109*(1), 75-90.

Gomez, P., Perea, M., & Ratcliff, R. (2007). A model of the Go/No-Go task. *Journal of Experimental Psychology-General, 136*(3), 389-413.

Gonzalez, R., & Nelson, T. O. (1996). Measuring ordinal association in situations that contain tied scores. *Psychological Bulletin, 119*(1), 159-165.

Goodman, L. A., & Kruskal, W. H. (1953). Measures of Association for Cross-Classifications. *Annals of Mathematical Statistics, 24*(1), 147-147.

Goodman, L. A., & Kruskal, W. H. (1954). Measures of Association for Cross Classifications. *Journal of the American Statistical Association, 49*(268), 732-764.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* Oxford, England: Wiley.

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology, 24*(3), 411-435.

Heath, R. A. (1984). Random-Walk and Accumulator Models of Psychophysical Discrimination - a Critical-Evaluation. *Perception, 13*(1), 57-65.

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2002). Quantile maximum likelihood estimation of response time distributions. *Psychonomic Bulletin & Review, 9*, 394-401.

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

Henmon, V. A. C. (1911). The relation of the time of a judgment to its accuracy. *Psychological Review, 18*, 186-201.

Irwin, F. W., Smith, W. A. S., & Mayfield, J. F. (1956). Tests of two theories of decision in an "expanded judgment" situation. *Journal of Experimental Psychology, 51*, 261-268.

Johnson, D. M. (1939). Confidence and speed in the two-category judgment. *Archives of Psychology, 34*, 1-53.

Johnson, J. G., & Busemeyer, J. R. (2005). A dynamic, stochastic, computational model of preference reversal phenomena. *Psychological Review, 112*(4), 841-861.

Juslin, P. (1994). The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items. *Organizational Behavior & Human Decision Processes, 57*(2), 226-246.

Juslin, P., & Olsson, H. (1997). Thurstonian and Brunswikian origins of uncertainty in judgment: A sampling model of confidence in sensory discrimination. *Psychological Review, 104*(2), 344-366.

Juslin, P., Winman, A., & Olsson, H. (2000). Naive empiricism and dogmatism in confidence research: A critical examination of the hard-easy effect. *Psychological Review, 107*(2), 384-396.

Juslin, P., Winman, A., & Persson, T. (1995). Can Overconfidence Be Used as an Indicator of Reconstructive Rather Than Retrieval-Processes. *Cognition, 54*(1), 99-130.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*, 773.

Keren, G. (1988). On the Ability of Monitoring Non-Veridical Perceptions and Uncertain Knowledge - Some Calibration Studies. *Acta Psychologica, 67*(2), 95-119.

Keynes, J. M. (1948). *A Treatise on Probability*. London: MacMillan and Co., Ltd.

Klayman, J., Soll, J. B., Gonzalez-Vallejo, C., & Barlas, S. (1999). Overconfidence: It depends on how, what, and whom you ask. *Organizational Behavior & Human Decision Processes, 79*(3), 216-247.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning & Memory, 6*(2), 107-118.

Laming, D. R. J. (1968). *Information Theory of Choice-Reaction Times*. New York, NY: Academic Press.

Lee, M. D., & Cummins, T. D. R. (2004). Evidence accumulation in decision making: Unifying the "take the best" and the "rational" models. *Psychonomic Bulletin & Review, 11*(2), 343-352.

Lichtenstein, S., & Fischhoff, B. (1980). Training for Calibration. *Organizational Behavior and Human Performance, 26*(2), 149-171.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306-334). Cambridge, England: Cambridge University Press.

Link, S. W. (1992). *The Wave Theory of Difference and Similarity*. Hiillsdale, N.J.: Lawrence Erlbaum Associates.

Link, S. W. (2003). C. S. Pierce, Confidence and Random Walk Theory. In B. B. E. Borg (Ed.), *Proceeding of the Nineteenth Annual Meeting of the International Society ofr Psychophysics*. Stockholm, Sweden: International Society for Psychophysics.

Link, S. W., & Heath, R. A. (1975). Sequential Theory of Psychological Discrimination. *Psychometrika, 40*(1), 77-105.

Luce, R. D. (1986). *Response Times: Their role in inferring elementary mental organization*. New York, NY: Oxford University Press.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guid*. New York, NY: Lawrence Erlbaum Associates, Inc.

McClelland, A. G. R., & Bolger, F. (1994). The calibration of subjective probabilities: Theories and models 1980-94. In G. Wright & P. Ayton (Eds.), *Subjective Probability* (pp. 453-482). Chichester: John Wiley & Sons.

Merkle, E. C., & Van Zandt, T. (2006). An application of the Poisson race model to confidence calibration. *Journal of Experimental Psychology-General, 135*(3), 391-408.

Murphy, A. H. (1973). A New Vector Partition of the Probability Score. *Journal of Applied Meteorology, 12*(4), 595-600.

Murphy, A. H., & Winkler, R. L. (1977). Reliability of subjective probability forecasts of precipitation and temperature. *The Journal of the Royal Statistical Society, 26*(1), 41-47.

Nelder, J. A., & Mead, R. (1965). A Simplex-Method for Function Minimization. *Computer Journal, 7*(4), 308-313.
Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95*(1), 109-133.
Nelson, T. O. (1987). The Goodman-Kruskal gamma coefficient as an alternative to signal-detection theory's measures of absollute-judgment accuracy. In E. Roskam & R. Suck (Eds.), *Progress in mathematical psychology*. New York, NY: Elsevier Science.
Nelson, T. O. (1996). Consciousness and metacognition. *American Psychologist, 51*(2), 102-116.
Nelson, T. O. (1997). The meta-level versus object-level distinction (and other issues) in formulations of metacognition. *American Psychologist, 52*(2), 179-180.
Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *The Psychology of Learning and Motivation, 26*, 125-141.
Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review, Vol 104(2)*, 266-300.
Pachella, R. G. (1974). The interpretation of reaction time in information-processing research. In.
Petrusic, W. M., & Baranski, J. V. (2003). Judging confidence influences decision processing in comparative judgments. *Psychonomic Bulletin & Review, 10*(1), 177-183.
Pierce, C. S. (1877). Illustrations of the logic of science. *The Popular Science Monthly, 12*, 1-15, 286-302, 604-615, 705-718.
Pierce, C. S., & Jastrow, J. (1884). On small differences of sensation. *Memoirs of the National Academy of Sciences, 3*, 73-83.
Pierrel, R., & Murray, C. S. (1963). Some Relationships between Comparative Judgment, Confidence, and Decision-Time in Weight-Lifting. *American Journal of Psychology, 76*(1), 28-&.
Pike, R. (1973). Response Latency Models for Signal-Detection. *Psychological Review, 80*(1), 53-68.
Raftery, A. E. (1995). Bayesian model selection in social research. In P. V. Marsden (Ed.), *Sociological methodology 1995* (pp. 111-196). Cambridge, MA: Blackwell.
Ratcliff, R. (1978). Theory of Memory Retrieval. *Psychological Review, 85*(2), 59-108.
Ratcliff, R., Gronlund, S. D., & Sheu, C. F. (1992). Testing Global Memory Models Using Roc Curves. *Psychological Review, 99*(3), 518-535.
Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science, 9*(5), 347-356.
Ratcliff, R., & Rouder, J. N. (2000). A diffusion model account of masking in two-choice letter identification. *Journal of Experimental Psychology-Human Perception and Performance, 26*(1), 127-140.
Ratcliff, R., & Smith, P. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological review, 111*(2), 333.
Ratcliff, R., & Starns, J. J., (2009). Modeling confidence and response time in recognition memory. *Psychological Review,* 116, 59-83.
Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review, 9*(3), 438-481.
Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review, 106*(2), 261-300.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review, 107*(2), 358-367.

Roe, R. M., Busemeyer, J. R., & Townsend, J. T. (2001). Multialternative decision field theory: A dynamic connectionst model of decision making. *Psychological Review, 108*(2), 370-392.

Rottenstreich, Y., & Tversky, A. (1997). Unpacking, repacking, and anchoring: Advances in support theory. *Psychological Review, 104*(2), 406-415.

Savage, L. J. (1954). *The Foundations of Statistics*. New York, NY: John Wiley & Sons.

Schouten, J. F., & Bekker, J. A. M. (1967). Reaction Time and Accuracy. *Acta Psychologica, 27*, 143-&.

Shadish, W. R., & Haddock, K. C. (1994). Combining estimates of effect size. In H. Cooper & L. V. Hedges (Eds.), *The Handbook of Research Synthesis*. New York: Russel Sage Foundation.

Shadlen, M. N., Kiani, R., Hanks, T., & Churchland, A. K. (2008). Neurobiology of Decision Making: An Intentional Framework. In C. Engel & W. Singer (Eds.), *Better Than Conscious? Decision Making, the Human Mind, and Implications For Institutions*. Cambridge, MA: MIT Press.

Shadlen, M. N., & Newsome, W. T. (2001). Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *Journal of Neurophysiology, 86*(1916-1936).

Sieck, W. R., Merkle, E. C., & Van Zandt, T. (2007). Option fixation: A cognitive contributor to overconfidence. *Organizational Behavior & Human Decision Processes, 103*, 68-83.

Smith, P. L. (1990). A note on the distribution of response time for a random walk with Gaussian increments. *Journal of Mathematical Psychology,* 34, 445-459.

Smith, P. L. (1995). Psychophysically Principled Models of Visual Simple Reaction-Time. *Psychological Review, 102*(3), 567-593.

Smith, P. L. (2000). Stochastic dynamic models of response time and accuracy: A foundational primer. *Journal of Mathematical Psychology, 44*(3), 408-463.

Squire, L. R., Wixted, J. T., & Clark, R. E. (2007). Recognition memory and the medial temporal lobe: A new perspective. *Nature Reviews Neuroscience, 8*(11), 872-883.

Stäel von Holstein, C., (1970). Measurement of subjective probability, *Acta Psychologica, 34, 146-159.*

Stone, M. (1960). Models for choice-reaction time. *Psychometrika, 25*(3), 251-260.

Suantak, L., Bolger, F., & Ferrell, W. R. (1996). The hard-easy effect in subjective probability calibration. *Organizational Behavior & Human Decision Processes, 67*(2), 201-221.

Svenson, O., & Maule, A. J. (1993). *Time Pressure and Stress in Human Judgment and Decision Making*. New York, NY: Plenum Press.

Swensson, R. G. (1972). Elusive Tradeoff - Speed Vs Accuracy in Visual Discrimination Tasks. *Perception & Psychophysics, 12*(1A), 16-&.

Swensson, R. G., & Edwards, W. (1971). Response strategies in a two-choice reaction task with a continuous cost for time. *Journal of Experimental Psychology*, 88, 67-81.

Thurstone, L. L., (1927). A law of comparative judgment. *Psychological Review*, 34, 273-286.

Tosoni, A., Galati, G., Romani, G. L., & Corbetta, M. (2008). Sensory-motoer mechanism in human parietal cortex underlie arbitrary visual decisions. *Nature Neuroscience, 11*, 1446-1453.

Townsend, J. T., & Ashby, F. G. (1983). *Stochastic modeling of elementary psychological proceses*. New York, NY: Cambridge University Press.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124-1131.

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review, 101*(4), 547-567.

Van Zandt, T. (2000a). How to fit a response time distribution. *Psychonomic Bulletin & Review, 7*(3), 424-465.

Van Zandt, T. (2000b). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 26*(3), 582-600.

Van Zandt, T., Colonius, H., & Proctor, R. W. (2000). A comparison of two response time models applied to perceptual matching. *Psychonomic Bulletin & Review, 7*(2), 208-256.

Van Zandt, T., & Maldonado-Molina, M. M. (2004). Response Reversals in Recognition Memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 30*, 1147-1166.

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

Vickers, D. (2001). Where does the balance of evidence lie with respect to confidence? In E. Sommerfeld, R. Kompass & T. Lachmann (Eds.), *Proceedings of the seventeenth annual meeting of the international society for psychophysics* (pp. 148-153). Lengerich: Pabst.

Vickers, D., Burt, J., Smith, P., & Brown, M. (1985). Experimental Paradigms Emphasizing State or Process Limitations .1. Effects on Speed Accuracy Tradeoffs. *Acta Psychologica, 59*(2), 129-161.

Vickers, D., & Packer, J. (1982). Effects of Alternating Set for Speed or Accuracy on Response-Time, Accuracy and Confidence in a Unidimensional Discrimination Task. *Acta Psychologica, 50*(2), 179-197.

Vickers, D., & Smith, P. (1985). Accumulator and random-walk models of psychophysical discrimination: A counter-evaluation. *Perception, 14*(4), 471-497.

Vickers, D., Smith, P., Burt, J., & Brown, M. (1985). Experimental Paradigms Emphasizing State or Process Limitations .2. Effects on Confidence. *Acta Psychologica, 59*(2), 163-193.

Volkman, J. (1934). The relation of the time of judgment to the certainty of judgment. *Psychological Bulletin, 31*, 672-673.

von Winterfeldt, D., & Edwards, W. (1986). *Decision Analysis and Behavioral Research*. New York, NY: Cambridge University Press.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review, 14*(5), 779-804.

Wagenmakers, E. J., van der Maas, H. L. J., & Grasman, R. P. P. P. (2007). An EZ-diffusion model for response time and accuracy. *Psychonomic Bulletin & Review, 14*(1), 3-22.

Wald, A. (1947). *Sequential Analysis*. New York, NY: Wiley.

Wald, A., & Wolfowitz, J. (1948). Optimum Character of the Sequential Probability Ratio Test. *Annals of Mathematical Statistics, 19*(3), 326-339.

Wallsten, T. S. (1996). An analysis of judgment research analyses. *Organizational Behavior & Human Decision Processes, 65*(3), 220-226.

Wallsten, T. S., Budescu, D. V., Erev, I., & Diederich, A. (1997). Evaluating and combining subjective probability estimates. *Journal of Behavioral Decision Making, 10*(3), 243-268.

Wallsten, T. S., Budescu, D. V., & Zwick, R. (1993). Comparing the calibration and coherence of numerical and verbal probability judgments. *Management Science, 39*(2), 176-190.

Wickelgren, W. A. (1977). Speed-Accuracy Tradeoff and Information-Processing Dynamics. *Acta Psychologica, 41*(1), 67-85.

Winkler, R. L., & Murphy, A. H. (1968). "Good" probability assessors. *Journal of Applied Meteorology, 7*(5), 751-758.

Winman, A., & Juslin, P. (1993). Calibration of Sensory and Cognitive Judgments - 2 Different Accounts. *Scandinavian Journal of Psychology, 34*(2), 135-148.

Yates, J. F. (1990). *Judgment and decision making*. Englewood Cliffs, NJ: Prentice Hall, Inc.

Yates, J. F., & Curley, S. P. (1985). Conditional distribution analyses of probabilisitc forecasts. *Journal of Forecasting, 4*, 61-73.

Yonelinas, a. P. (1994). Receiver-Operating Characteristics in Recognition Memory - Evidence for a Dual-Process Model. *Journal of Experimental Psychology-Learning Memory and Cognition, 20*(6), 1341-1354.

Author Note

Timothy J. Pleskac, Department of Psychology, Michigan State University. 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. We thank Jim Townsend, Thomas Wallsten, and Avi Wershbale, for their input on this work. 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 in Memphis, TN; the 2007 Annual Meeting for the Society for Mathematical Psychology in Memphis, TN: the 2008 Annual Meeting for the Society for Mathematical Psychology in Washington D.C.; and the 2008 Annual Meeting for the Society for Judgment and Decision Making;

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 δ, starting point z, and choice threshold θ, and drift coefficient σ, the probability of choosing alternative *A*, R_A for a Wiener process is

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

The probability of incorrectly choosing R_A when presented with SB, $P(R_A|S_B)$, can be found by replacing δ with –δ in Equation A 1. 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 θ and the judge responds R_A given stimulus S_A is

Equation A 2

$$g(t_D|R_A, S_A) = \frac{1}{P(R_A|S_A)}\pi\left(\frac{2\theta}{\sigma}\right)^{-2} e^{\delta(\theta - z)/\sigma^2} \sum_{j=1,\infty} j\sin\left(\frac{j\pi(\theta - z)}{2\theta}\right) e^{-\frac{dt}{2}(\delta^2/\sigma^2 + (\pi j\sigma/2\theta)^2)}.$$

The cumulative distribution function is

Equation A 3

$$G(t_D|R_A, S_A) = P(R_A|S_A) - \pi\left(\frac{2\theta}{\sigma}\right)^{-2} e^{\delta(\theta - z)/\sigma^2} \sum_{j=1,\infty} \frac{2j\sin\left(\frac{j\pi(\theta - z)}{2\theta}\right) e^{-\frac{dt}{2}(\delta^2/\sigma^2 + (\pi j\sigma/2\theta)^2)}}{\delta^2/\sigma^2 + (\pi j\sigma/2\theta)^2}.$$

The expressions for the pdf and cdf of the finishing times when stimulus S_B is present can be found by replacing δ with –δ 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. The distributions for confidence ratings are given in the text Equation 16.

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 *v* and a standard deviation η, $f(\delta) \sim N(v, \eta)$. The value of the starting point was assumed to be uniformally distributed with a range s_z, u(z) ~ 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 δ and z.

Appendix B

Choice probabilities, mean decision time statistics, and decision time and confidence distributions for Poisson models.

The reader is referred to Townsend and Ashby (1983) and/or Van Zandt, Colonius, and Proctor (2000) for a full development of the model as well as Van Zandt (2000b) for the derivation of the confidence distributions.

According to the Poisson race model, if presented with stimulus S_A, assuming accrual rates v_A and v_B, and choice thresholds K_A and K_B, the probability of choosing alternative A is

Equation B 1 $$P(R_A|S_A) = \sum_{j=0}^{K_B-1} \binom{K_A + j - 1}{j} \left(\frac{v_A}{v_A+v_B}\right)^{K_A} \left(\frac{v_B}{v_A+v_B}\right)^{j}.$$

The probability of choosing alternative B is found by switching K_B for K_A and K_A for K_B as well as switching v_B for v_A and v_A for v_B.

Assuming the two counters receive unit counts of evidence at exponentially distributed inter-arrival times, then the decision time t_D is gamma distributed (Townsend & Ashby, 1983). The race model probability density function for t_D when responding A is then

Equation B 2 $$g(t_D|R_A,S_A) = \frac{1}{P(R_A|S_A)} \frac{(v_A dt)^{K_A-1} v_A e^{-v_A dt}}{(K_A-1)!} \sum_{j=0}^{K_B-1} \frac{v_B dt e^{-v_B dt}}{j!}.$$

Integrating $g(\boldsymbol{t_D}|R_A,S_A)$ produces the cumulative distribution function of t_D when responding A, Equation B 3

$$G(t_D|R_A,S_A) =$$

$$1 - \frac{e^{-dt(v_A+v_B)}}{P(R_A|S_A)} \left(\frac{v_A}{(v_A+v_B)}\right)^{K_A} \sum_{i=0}^{K_B-1} \binom{K_A + i - 1}{i} \left(\frac{v_B}{(v_A+v_B)}\right)^{i} \sum_{j=0}^{K_A+i-1} \frac{[dt(v_A+v_B)]^j}{j!}.$$

The pdf and cdf of t_D when choosing alternative B is found by switching K_B for K_A and K_A for K_B, switching v_B for v_A and v_A for v_B, and substituting changing the choice probability.

Assuming Vicker's (1979) balance of evidence hypothesis, Van Zandt (2000) derived the predicted distribution of confidence ratings for the no-choice full scale method of confidence ratings.

In terms of a race model for two-alternative forced choice questions, the balance of evidence hypothesis states that one's confidence is the difference between the two counters. Therefore, when response alternative A is given (R_A), confidence in the Poisson race model is,

Equation B 4 $$C = X_A(\boldsymbol{t_D}) - X_B(\boldsymbol{t_D}) = K_A - X_B(\boldsymbol{t_D}).$$

The random variable C can take on any value between $1 \leq c \leq K_A$. The probability of C taking any of these values can be found by noticing that C is a linear function of the total counts of evidence that have arrived across the two counters

Equation B 5 $$C = 2K_A - Y,$$

where,

Equation B 6 $$Y = X_A(t_D) + X_B(t_D) = K_A + X_B(t_D).$$

This relationship between evidence counts and confidence illustrates that confidence in the Poisson model using the balance of evidence hypothesis is in many ways a more elaborate version of the early time based hypotheses of confidence where confidence is an inverse function of decision time (Audley, 1960; Volkman, 1934).

As in the choice probabilities, the random variable Y has a negative binomial distribution where one is waiting for K_A evidence counts on counter A in a sequence of Y Bernoulli trials. So that when the counter for response alternative A wins,

Equation B 7 $$P(Y = y \cap A\ wins) = \binom{y-1}{K_A-1} p^{K_A}(1-p)^{y-K_A}.$$

Where $= \frac{v_A}{v_A+v_B}$. Using Equation B 5, we can substitute $y = 2K_A - c$ into Equation C 7 and divide by the choice probability $P(R_A|S_A)$ to find the distribution over possible confidence values,

Equation B 8 $$P(C = c|R_A, S_A) = \binom{2K_A - c - 1}{K_A - 1} \left(\frac{v_A}{v_A+v_B}\right)^{K_A} \left(\frac{v_B}{v_A+v_B}\right)^{K_A-c}.$$

The derivations when relative balance of evidence are found in a similar manner (see Equation 17). The reader is referred to Merckle and Van Zandt (2006) for the relative balance of evidence derivations.

A problem with the balance of evidence hypothesis is that the discrete nature of the confidence responses limits the applicability of the Poisson race model. The balance of evidence mapping hypothesis is one possible solution to this problem. Under this hypothesis we retained the discrete counts, but we posited that C is perturbed with noise forming a larger population of graded levels of covert confidence, Ω. This was modeled by assuming a Gaussian density was

centered over *C* so that height of the density is a weighted average of all the normal densities centered over each value of C, which is akin to using a Gaussian Kernel to estimate a density function in statistics (Van Zandt, 2000a). See Equation 18 for more details.

Appendix C

2DSD with trial variability Parameter Estimates for Each participant in Each Task

The table below lists the parameter estimates for each participant in the two tasks. The parameter estimates were found using the adapted QML method for confidence ratings.

--

Insert Table C 1

Appendix D

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

This appendix describes the Makov chain approximation of the 2DSD model using the marker hypothesis of confidence. 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.

In the model the choice stage works as follows. The state space of evidence L ranges from the lower choice threshold –θ to the upper threshold θ as a function of step size Δ. Consequently, L can be expressed as a function of step size

Equation D1

$$L = \begin{matrix} \{-\mathrm{k}\Delta & -(\mathrm{k}-1)\Delta, & \cdots & -\Delta, & 0, & \Delta, & \cdots & (\mathrm{k}-1)\Delta, & \mathrm{k}\Delta\ \} \\ 1 & 2 & \cdots & (m-1)/2 & & \cdots & m-1 & m \end{matrix}$$

Where $\theta = \mathrm{k}\Delta$ and $-\theta = -\mathrm{k}\Delta$. 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 $1 < i < m$,

Equation D 2

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

Where the drift and diffusion coefficients are δ and σ^2 respectively. The parameter ρ is the time interval that passes with each sampled piece of interval. As ρ approach zero and setting $\Delta = \alpha\sigma\sqrt{\rho}$ the random walk will converge to a Weiner diffusion process which has a continuous time set and continuous state space. 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 $P = \|p_{i,j}\|$ is presented in its canonical form:

Equation D 3

$$P = \left[\begin{array}{c|c} P_1 & 0 \\ \hline R & Q \end{array}\right]$$

	1	m	2	3	…	m-2	m-1
1	1	0	0	0	…	0	0
m	0	1	0	0	…	0	0
2	$p_{2,1}$	0	$p_{2,2}$	$p_{2,3}$	…	0	0
3	0	0	$p_{3,2}$	$p_{3,3}$	…	0	0
4	0	0	0	$p_{4,3}$	…	0	0
	⋮	⋮	⋮	⋮	⋯	⋮	⋮
m-3	0	0	0	0	…	$p_{m-3,m-2}$	0
m-2	0	0	0	0	…	$p_{m-2,m-2}$	$p_{m-2,m-1}$
m-1	0	$p_{m-1,m}$	0	0	…	$p_{m-1,m-2}$	$p_{m-1,m-1}$

With P_1 being a 2 x 2 matrix with two absorbing states, one for each choice alternative. Q is an (m-2) x (m-2) matrix that contains the transition probabilities $p_{i,j}$ (see Equation E 2). R is an (m-2) x 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*, *Pr*(*A*), and the probability of choosing option *B*, *Pr*(*B*)) are

Equation D 4 $$[\Pr(A), \Pr(B)] = Z \cdot (I - Q)^{-1} \cdot R,$$

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) x (m-2). The mean decision time conditional on each choice is

Equation D 5 $$[\mathrm{E}(T_D|B), \mathrm{E}(T_D|A)] = Z \cdot (I - Q)^{-2} \cdot R./[\Pr(A), \Pr(B)].$$

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 c_i are placed along the evidence state space representing the different confidence ratings (.50...1.00), one for each rating. For the confidence ratings below certainty (< 1.00), 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 best performing model assumed that the state associated with the confidence rating 1.00 is an absorbing boundary. Thus, there was one absorbing boundary in the second stage. Moreover, we assume that evidence never drops below a certain level so that the evidence accumulation process is reflected on a lower boundary. We placed the lower boundary far away (VΔ) from the starting point in the first stage so that this boundary had essentially no effect and the process approximated a process with an unrestricted boundary.

In what follows we will describe the Markov chain approximation of this second stage assuming the upper boundary θ 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

$$L_A = \{-V\Delta, \quad \cdots \quad \Delta, \quad 0, \quad \Delta, \quad 2\Delta \quad \cdots \quad k\Delta\} \cup \{c_{.50}, c_{.60}, c_{.70}, c_{.80}, c_{.90}\}.$$

Where in this case $c_{1.00} = k\Delta$. The transition probability matrix is similar to that given in Equation E 3 except for three changes. A new transition matrix P^A is formed that is ((m+5) x (m+5)) where $m = k\Delta + V\Delta + 1$. The values in the transition matrix for $p_{i,j}$ are given by Equation D2. The transition matrix has an additional 5 rows and 5 columns to account for the possibility of giving a confidence rating (conf = .50, .60…, .90) associated with one of the five confidence markers c_{conf}. P^A in its canonical form is:

$P^A =$ $\begin{bmatrix} P_1{}^A & 0 \\ R^A & Q^A \end{bmatrix}$ =

	m+1	…	m+5	m	1	2	…	$m_{.50}^*$	…	$m_{.90}^*$	…	m-2	m-1
m+1	1	…	0	0	0	0	…	0	…	0	…	0	0
⋮	⋮	⋯	⋮	⋮	⋮	⋮	⋯	⋮	⋯	⋮	⋯	⋮	⋮
m+5	0	…	1	0	0	0	…	0	…	0	…	0	0
m	0	…	0	1	0	0	…	0	…	0	…	0	0
1	0	…	0	0	$p_{1,1}$	$p_{1,2}$	…	0	…	0	…	0	0
2	0	…	0	0	$p_{2,1}$	$p_{2,2}$	…	0	…	0	…	0	0
⋮	⋮	⋯	⋮	⋮	⋮	⋮	⋯	⋮	⋯	⋮	⋯	⋮	⋮
$m_{.50}^*$	$w_{.50}$	…	0	0	0	0	…	$(1-w_{.50})p_{m_{.50}^* m_{.50}^*}$	…	0	…	0	0
⋮	⋮	⋯	⋮	⋮	⋮	⋮	⋯	⋮	⋯	⋮	⋯	⋮	⋮
$m_{.90}^*$	0	…	$w_{.90}$	0	0	0	…	0	…	$(1-w_{.90})p_{m_{.90}^* m_{.90}^*}$	…	0	0
⋮	⋮	⋯	⋮	⋮	⋮	⋮	⋯	⋮	⋯	⋮	⋯	⋮	⋮
m-2	0	…	0	0	0	0	…	0	…	0	…	$p_{m-2,m-2}$	$p_{m-2,m-1}$
m-1	0	…	0	$p_{m-1,m}$	0	0	…	0	…	0	…	$p_{m-1,m-2}$	$p_{m-1,m-1}$

In general it takes a similar form as the probability transition matrix P used during the choice stage. The changes are as follows. First, there is a reflecting boundary at the bottom evidence state $V\Delta$ so P(1,1) is in the transition matrix Q^A. To allow for the possibility of exiting the evidence accumulation process and giving a confidence rating the evidence state at row m_{conf}^* corresponding to the index of P^A associated with the confidence marker c_{conf} is multiplied by (1-w_{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 $P^{A}_{(m+1),(m+1)}, \ldots, P^{A}_{(m+5),(m+5)} = 1$ (see P_1^A). The last five columns of P^A contains all zeros except for the row corresponding to the absorbtion state for each confidence rating and the row corresponding to the confidence marker, m^{*}_{conf}, which is set to w_{conf} (see R^A).

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

Equation D 6 $$[\Pr(.50), \Pr(.60), \ldots \Pr(1.00)] = Z^A \cdot (I^A - Q^A)^{-1} \cdot R^A,$$

where ZA is a m-1 vector denoting the initial starting position of the process, the location of the choice threshold from the choice stage. The expected inter-judgment time is

Equation D 5

$$[\mathrm{E}(\tau|.50, A), \mathrm{E}(\tau|.60, A), \ldots \mathrm{E}(\tau|1.00, A)] = Z \cdot (I - Q)^{-2} \cdot R./[\Pr(.50), \Pr(.60), \ldots \Pr(1.00)].$$

In total the model, has the same parameters as shown in Table 2 except there are 6 confidence markers (instead of 5 confidence criteria) and 5 probabilities of exiting w_{conf}. To evaluate the model we fit the model to the accuracy and speed conditions of both the line length and city population tasks. For each condition the model was fit to the proportion of correct choice, mean decision time in correct and incorrect choices, the proportion of confidence ratings for correct and incorrect choices, and the mean inter-judgment times for each confidence rating conditional on correct and incorrect choices. Between the speed and accuracy conditions this produced 50 free data points in each task. We fit the model with least square methods assuming unit weighting and choice thresholds θ, w_{conf}, and the confidence marker $c_{1.00}$ were free to vary between the speed and accuracy conditions for a total of 17 free parameters for each task.

Footnotes

[1.] A third possible model is Juslin and Olsson's (1997) sampling model of sensory discrimination. Vickers and Pietsch (2001), however, have shown this model makes several counterintuitive and unsupported predictions. For example, the model is severely limited in its ability to correctly predict the speed-accuracy trade off.

[2.] Ratcliff's (Ratcliff, 1978; Ratcliff & Smith, 2004) formalization of a diffusion model places the lower threshold θ_B at the 0 point and places an unbiased starting point at the half-way point between the upper and lower thresholds.

[3.] 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.

[4.] This unbridled growth of accuracy is also seen as unrealistic aspect of random walk/diffusion models in general. 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) (cf. Bogacz et al., 2006; Busemeyer & Townsend, 1993) do not have this assumption.

[5.] We are grateful for Steve Link for pointing out the work of Pierce and Jastrow.

[6.] 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).

[7.] After considerable work on this model we found that Van Zandt and Maldonado-Molina (2004) made a similar hypothesis, but dismissed it as implausible. We are going to see how far this hypothesis gets us.

[8.] Vickers (1979) originally expressed this hypothesis in terms of his accumulator model for two-category discrimination which is a discrete time sequential sampling model with an absolute stopping rule. That is unlike the Poisson race model evidence is continuous, but time is discrete.

[9.] A second advantage of the QML method is that it is more robust to outliers in terms of decision times than maximum likelihood methods (Heathcote, et al., 2002).

[10.] 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.

[11.] 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.

[12.] Table 6 shows that there was also a relationship between inter-judgment time and difficulty, but our analyses showed this played little role in the hard-easy effect in this dataset.

[13.] 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).

Table 1. The Eight Empirical Hurdles a Model of Cognitive Performance Must Account For

	Hurdle	Description	References
1	Speed/Accuracy Tradeoff	Speed/Accuracy trade off where decision time and choice accuracy are inversely related such that the judge can trade accuracy for speed.	(Garrett, 1922; D. M. Johnson, 1939; Pachella, 1974; Schouten & Bekker, 1967; Wickelgren, 1977)
2	Positive relationship between confidence and stimulus difficulty	Confidence is a monotonically decreasing function of the difficulty of the stimuli.	(Baranski & Petrusic, 1998; Festinger, 1943; Garrett, 1922; D. M. Johnson, 1939; Pierce & Jastrow, 1884; Pierrel & Murray, 1963; Vickers, 1979)
3	Resolution of confidence	Choice accuracy and confidence are monotonically related even after controlling for the difficulty of the stimuli	(Ariely et al., 2000; Baranski & Petrusic, 1998; Dougherty, 2001; Garrett, 1922; D. M. Johnson, 1939; Nelson & Narens, 1990; Vickers, 1979)
4	Negative relationship between confidence and decision time	During optional stopping tasks there is a monotonically decreasing relationship between the decision time and confidence where judges are more confident in fast decisions.	(Baranski & Petrusic, 1998; Festinger, 1943; D. M. Johnson, 1939; Vickers & Packer, 1982)
5	Positive relationship between confidence and decision time	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), there is a monotonically increasing relationship between	(Irwin, Smith, & Mayfield, 1956; Vickers & Packer, 1982; Vickers, Smith et al., 1985)

		confidence and decision time where participants are on average more confident in conditions when they are take more time to make a choice.	
6	Slow Errors	For difficult conditions, particularly when accuracy is emphasized, mean decision times for incorrect choices are slower than mean decision times for correct choices.	(Luce, 1986; Ratcliff & Rouder, 1998; Swensson, 1972; Townsend & Ashby, 1983; Vickers, 1979)
7	Fast Errors	For easy conditions, particularly when speed is emphasized, mean decision times for incorrect choices are faster than mean decision times for correct choices.	(Ratcliff & Rouder, 1998; Swenson & Edwards, 1971; Townsend & Ashby, 1983)
8	Increased resolution in confidence with time pressure	When under time pressure at choice, there is an increase in the resolution of confidence judgments.	(Current Paper; Baranski & Petrusic, 1994)

Table 2. Parameters of the 2 Stage Dynamic Signal Detection Model of Confidence

Parameter	Description
δ	*Drift rate.* Controls the average rate of accumulation across trials and indexes the average strength or quality of the evidence judges are able to accumulate
σ^2	*Drift coefficient.* Responsible for the within-trial random fluctuations. It is unidentifiable and is set to .1
θ_A, θ_B	Choice threshold. Determines the quantity of information judges accumulate before selecting a choice. Controls the speed-accuracy tradeoff.
z	*Start point.* Determine the point in the evidence space where judges begin accumulating evidence.
t_{ED}	*Mean non-decision time*. Observed decision time is a function of the non decision time and decision time predicted by the model, $t_D' = t_E + t_D$
$c_{choice,k}$	*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.
τ	*Inter-judgment time.* A parameter indexing the time between when a decision is made and a confidence rating is entered.
t_{EJ}	*Mean non-judgment time*. Observed inter-judgment time is a function of the non judgment time and inter-judgment time used in the model, $\tau' = t_{EJ} + \tau$

Table 3. Parameters of the Poisson Race Models of Confidence

Parameter	Description
v_A	*Accrual rate.* Information accrual rate for the correct choice
r	*Total accrual rate.* Sum of information accrual rates $v_A + v_B$.
K_A, K_B	*Response criteria.* Determines the quantity of information accumulated. Controls the speed/accuracy tradeoff.
t_{ED}	*Mean non-decision.*Observed decision time is a function of the non decision time and decision time predicted by the model, $t_D' = t_{ED} + t_D$
Parameters for Confidence mapping hypothesis.	
h	*Bandwidth parameter.* Controls the width of the normal distributions centered over discrete confidence values Y_i
c_i	*Confidence criteria.* Section the covert confidence variable Ω to map a confidence rating to it.

Table 4 Mean, Standard Error, and Standard Deviation Between Participants of Button Press Time in Seconds for Each Confidence Rating

Confidence						
Button	.50	.60	.70	.80	.90	1.00
Mean	0.316	0.311	0.327	0.268	0.317	0.299
SE	0.007	0.007	0.007	0.006	0.007	0.006
Std_{btwn}	0.062	0.072	0.059	0.076	0.071	0.067

The statistics were calculated after removing trials that were greater than 3 standard deviations from the untrimmed mean.

Table 5. Proportion Correct, Average Decision Time, Average Confidence Rating, and Average Inter-Judgment Time for Each Participant.

		Par.	1	2	3	4	5	6	Ave.
Line	Prop. Correct	Speed	.82*	.80*	.73*	.76*	.82*	.78*	.79*
Length		Acc.	.87	.82	.86	.88	.85	.87	.86
	Decision Time (s)	Speed	0.55 (0.19)*	0.52 (0.13)*	0.45 (0.11)*	0.54 (0.1)*	0.51 (0.13)*	0.55 (0.1)*	0.52*
		Acc.	0.73 (0.47)	0.74 (0.39)	0.91 (0.71)	1.75 (1.49)	0.73 (0.46)	1.59 (1.32)	0.80
	Confidence	Speed	.82 (.22)*	.85 (.16)	.87 (.17)*	.89 (.15)*	.97 (.12)*	.76 (.18)*	.87*
		Acc.	.89 (.18)	.84 (.17)	.94 (.12)	.96 (.08)	.99 (.07)	.84 (.17)	.95
	Inter-judgment time	Speed	0.55 (0.50)	0.63 (0.57)	0.59 (0.38)	1.06 (0.84)	0.34 (0.31)	1.04 (0.64)	0.56
	(s)	Acc.	0.55 (0.53)	0.54 (0.36)*	0.50 (0.3)*	0.43 (0.36)*	0.28 (0.19)*	0.71 (0.47)*	0.42*
City	Prop. Correct	Speed	.60*	.69*	.67*	.68*	.67*	.66	.66*
Population		Acc.	.64	.75	.78	.78	.73	.68	.73
	Decision Time (s)	Speed	0.84 (0.28)*	1.08 (0.17)*	0.91 (0.22)*	1.05 (0.21)*	0.97 (0.27)*	1.11 (0.2)*	1.02*
		Acc.	1.19 (0.72)	1.6 (0.65)	2.33 (1.51)	2.71 (1.53)	2.53 (1.61)	2.46 (1.22)	1.74
	Confidence	Speed	.58 (.16)*	.78 (.18)*	.81 (.15)*	.83 (.15)*	.83 (.20)*	.75 (.17)*	.76*
		Acc.	.60 (.17)	.80 (.17)	.85 (.12)	.89 (.10)	.87 (.17)	.78 (.16)	.82
	Inter-judgment time	Speed	0.60 (0.56)	0.88 (0.45)	0.68 (0.52)	1.27 (0.78)	1.23 (0.81)	1.72 (0.8)	0.93
	(s)	Acc.	0.56 (0.5)*	0.58 (0.25)*	0.33 (0.21)*	0.35 (0.18)*	0.66 (0.5)*	1.35 (0.74)*	0.44*

Values in parentheses are standard deviations. * indicates the condition (speed or accuracy) in which a z test revealed the relevant statistic was smaller using an alpha value of .05 (one-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.

Table 6. Average (Between Par. Std) Goodman and Kruskal γ Correlation Coefficient Across All 6 Participants for the Line Length and City Population Discrimination Tasks During Accuracy (below diagonal) and Speed (above diagonal).

Line Length	Obj. Difference Between Line Lengths	Accuracy	Decision Time	Confidence	Inter-judgment time
Obj. Difference Between Line Lengths	Acc \ Speed	.50 (.06)	-.13 (.04)	.36 (.08)	-.16 (.07)
Accuracy	.68 (.05)	Acc \ Speed	-.12 (.11)	.76 (.08)	-.31 (.1)
Decision Time	-.27 (.08)	-.27 (.09)	Acc \ Speed	-.23 (.18)	.14 (.06)
Confidence	.44 (.14)	.67 (.11)	-.47 (.29)	Acc \ Speed	-.54 (.25)
Inter-judgment time	-.12 (.1)	-.19 (.12)	.24 (.12)	-.46 (.26)	Acc \ Speed
City Population	**Obj. Difference Between City Populations**	**Accuracy**	**Decision Time**	**Confidence**	**Inter-judgment time**
Obj. Difference Between City Populations	Acc \ Speed	.26 (.08)	-.06 (.05)	.18 (.04)	-.07 (.05)
Accuracy	.34 (.08)	Acc \ Speed	-.10 (.12)	.55 (.09)	-.16 (.12)
Decision Time	-.09 (.04)	-.17 (.08)	Acc \ Speed	-.16 (.16)	.08 (.03)
Confidence	.20 (.05)	.43 (.06)	-.37 (.15)	Acc \ Speed	-.32 (.26)
Inter-judgment time	-.03 (.05)	-.08 (.08)	.12 (.05)	-.19 (.22)	Acc \ Speed

The average Goodman and Kruskal γ correlation coefficients were calculated by weighting each subject's respective coefficient by the inverse of the variance of the γ (see Goodman & Kruskal, 1963). The values in the parentheses are an estimate of the between participant standard deviation. Due to the high number of observations per participant the standard errors for each estimate in the table above is ≤ .01. Thus, each γ in the table above is significant at an alpha level of .05.

Table 7. Measures of Slope and Scatter for Each Participant in the Speed and Accuracy Conditions of Each Task.

		Par.	1	2	3	4	5	6	Ave
Line Length	Slope	Speed	.22 (.01)	.17 (.01)	.21 (.01)	.20 (.01)	.14 (.01)	.21 (.01)	.19 (0)
		Accuracy	.19 (.01)*	.17 (.01)	.12 (.01)*	.07 (0)*	.05 (0)*	.16 (.01)*	.09 (0)*
	Scatter	Speed	.040 (.002)	.022 (.001)	.019 (.001)	.016 (.001)	.013 (.001)	.025 (.001)	.021 (.001)
		Accuracy	.030 (.002)*	.023 (.001)	.013 (.001)*	.006 (.001)*	.005 (.001)*	.024 (.0009)	.011 (.001)*
City	Slope	Speed	.08 (.01)	.13 (.01)	.14 (.01)	.15 (.01)	.14 (.01)	.10 (.01)	.12 (.00)
Population		Accuracy	.08 (.01)	.10 (.01)	.08 (.01)*	.07 (.01)*	.10 (.01)*	.08 (.01)*	.08 (.00)*
	Scatter	Speed	.022 (.001)	.026 (.001)	.017 (.001)	.017 (.001)	.037 (.002)	.027 (.001)	.021 (.001)
		Accuracy	.027 (.001)	.026 (.001)	.012 (.001)*	.009 (.001)*	.028 (.002)*	.024 (.001)*	.014 (.001)*

Values in parentheses are standard errors. * indicates the condition (speed or accuracy) in which a z test revealed the relevant statistic was smaller using an alpha value of .05 (one tailed).

Table 8. Description of the Five Model Fits Reported and their Parameter Constraints

Model	No. Parameters Line / City	Parameter Constraints
Baseline	200 / 240	The observed marginal relative frequencies are used as the predicted probabilities for both the decision time categories and confidence ratings in the QML method.
2DSD	14 / 15	There was no bias in the starting point ($z = 0/ \theta_A = \theta_B = \theta$). Only the drift rate δ was allowed to vary between levels of difficulty. The choice threshold was allowed to vary between speed and accuracy manipulations. Confidence criteria were symmetrical around the starting point z for corrects and incorrects and were held fixed across all conditions. All remaining parameters were also held constant across conditions.
2DSD with Trial Variability	16 / 17	The model fitting was identical in most cases to the 2DSD (above). Trial variability in the drift rate δ was modeled with a normal distribution with a mean ν and standard deviation η. Trial variability in the start point was modeled with a uniform distribution centered over $z = 0$ with a range s_z. See Appendix A for more details.
Poisson Race Model – Relative Balance of Evidence Hypothesis	9 / 10	There was no bias in the choice thresholds ($K_A = K_B = K$). The accumulation rate for the correct choice (v_A) was allowed to vary between levels of difficulty. All other parameters were held fixed across conditions.
Poisson Race Model – Mapping Balance of Evidence Hypothesis	15 / 16	There was no bias in the choice thresholds ($K_A = K_B = K$). The accumulation rate for the correct choice (v_A) was allowed to vary between levels of difficulty. All other parameters were held fixed across conditions.

Table 9. Bayesian Information Criterion (BIC) Values for Each Model and Participant in the Line Length Discrimination Task.

Par.	No. Obs.	Baseline (200)	2DSD (14)	Poisson Race - RBH (9)	Poisson Race - Map (15)
1	5,705	72,907.58	35,886.82*	58,858.58	37,506.30
2	5,750	81,108.04	40,046.29*	46,404.61	41,994.23
3	5,546	71,047.82	35,375.61*	54,869.66	39,075.04
4	5,707	69,407.31	35,517.41*	63,109.73	41,492.69
5	5,718	53,401.67	27,091.25*	38,839.41	28,787.16
6	5,756	82,330.07	42,176.47*	53,425.07	48,425.16

The values in the parentheses are the number of free parameters associated with each model in fitting correct and incorrect decision time and confidence distributions for corrects and incorrects for 10 different conditions (speed vs accuracy x 5 levels of difficulty). *indicates the best fitting model according to the BIC.

Table 10. Bayesian Information Criterion (BIC) Values for Each Model and Participant in the City Population Discrimination Task.

Par	No. Obs.	Baseline (240)	2DSD (15)	Poisson Race - RBH (10)	Poisson Race - Map (16)
1	3,978	45,460.37	27,689.50*	33,158.94	28,226.92
2	3,999	54,236.52	32,482.25*	38,097.92	33,872.96
3	3,824	48,522.03	30,174.30*	37,924.07	32,063.94
4	3,961	49,449.36	30,294.86*	40,621.04	32,492.42
5	3,225	40,449.26	24,284.11*	34,824.58	25,518.53
6	3,966	53,910.90	33,280.52*	38,726.16	34,950.69

The values in the parentheses are the number of free parameters associated with each model in fitting correct and incorrect decision time and confidence distributions for corrects and incorrects for 12 different conditions (speed vs accuracy x 6 levels of difficulty). *indicates the best fitting model according to the BIC.

Table 11. Bayesian Information Criterion (BIC) Values for the 2DSD Model with and without Trial-by-Trial Variability for Each Participant

	Line Length		City Population	
	2DSD	2DSD w Trial Variability	2DSD	2DSD w Trial Variability
1	35,886.82	35,871.46*	27,689.50	27,676.53*
2	40,046.29	40,036.95*	32,482.25*	32,486.62
3	35,375.61	35,369.75*	30,174.30	30,161.35*
4	35,517.41	35,491.60*	30,294.86	30,118.97*
5	27,091.25	27,035.69*	24,284.11	24,231.79*
6	42,176.47	41,989.86*	33,280.52	32,995.55*

* indicates best fitting model according to BIC

Table 12. The Average Goodman and Kruskal γ Between Confidence and Decision Time Holding Difficulty Constant for the Line Length Discrimination Task

	Participant	1	2	3	4	6	Ave.
Speed	Correct	-0.28	-0.35	0.00	-0.22	-0.31	-0.26
	Incorrect	-0.13	-0.31	0.09	0.08	-0.15	-0.06
Accuracy	Correct	-0.37	-0.33	-0.31	-0.56	-0.55	-0.46
	Incorrect	-0.25	-0.25	-0.04	-0.29	-0.30	-0.18

Participant 5 used primarily the .50, .90 and 1.00 confidence ratings during the line discrimination task and was thus excluded from these calculations.

Table 13. DI', Bias, and Brier Scores Across Participants in the Speed and Accuracy Conditions of Each Task.

		Line Length			City Population		
		Mean	SE	Std_{btwn}	Mean	SE	Std_{btwn}
DI'	Speed	1.30	0.02	0.19	0.78	0.02	0.23
	Accuracy	1.02*	0.03	0.12	0.59*	0.02	0.09
Bias	Speed	0.09	0.01	0.07	0.1	0.02	0.06
	Accuracy	0.08	0.01	0.06	0.08	0.01	0.07
Brier	Speed	0.140	0.0002	0.012	0.214	0.0002	0.014
	Accuracy	0.113*	0.0002	0.014	0.203*	0.0002	0.025

Brier score standard errors were estimated with a bootstrap method. * indicates a significantly predicted lower value according to a z test using an alpha value of .05 (one tail).

Table C 1. Parameter estimates from the 2DSD model with trial variability in the drift rate and starting point.

	Line Length						City Population					
Par	1	2	3	4	5	6	1	2	3	4	5	6
ν_1	0.039	0.030	0.021	0.018	0.036	0.022	0.001	0.014	0.017	0.012	0.009	0.013
ν_2	0.095	0.087	0.062	0.048	0.108	0.086	0.020	0.026	0.035	0.028	0.032	0.023
ν_3	0.172	0.166	0.119	0.107	0.194	0.147	0.025	0.045	0.051	0.050	0.041	0.042
ν_4	0.231	0.223	0.171	0.157	0.244	0.238	0.043	0.061	0.063	0.066	0.057	0.060
ν_5	0.297	0.326	0.235	0.233	0.304	0.347	0.049	0.094	0.077	0.096	0.074	0.093
ν_6							0.085	0.149	0.122	0.140	0.134	0.142
η	0.081	0.175	0.045	0.063	0.098	0.171	0.010	0.054	0.037	0.046	0.071	0.125
θ_{speed}	0.058	0.047	0.044	0.042	0.053	0.051	0.066	0.061	0.081	0.063	0.073	0.068
θ_{acc}	0.077	0.076	0.089	0.140	0.080	0.141	0.088	0.098	0.146	0.162	0.159	0.179
s_z	0.053	0.045	0.033	0.001	0.043	0.001	0.001	0.001	0.039	0.000	0.000	0.025
et	0.324	0.375	0.292	0.400	0.324	0.395	0.436	0.772	0.447	0.723	0.546	0.775
ejt	0.341	0.001	0.160	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.026
c_1	0.044	-0.098	-0.060	-0.107	-0.035	-0.030	0.120	-0.026	-0.038	-0.126	0.013	-0.123
c_2	0.062	-0.013	-0.005	-0.042	-0.034	0.084	0.158	0.040	0.031	-0.029	0.053	0.045
c_3	0.067	0.075	0.036	0.004	-0.034	0.174	0.172	0.090	0.084	0.065	0.088	0.161
c_4	0.071	0.133	0.051	0.050	-0.034	0.266	0.181	0.132	0.159	0.148	0.119	0.272
c_5	0.083	0.177	0.080	0.152	-0.030	0.353	0.186	0.171	0.191	0.205	0.140	0.388

Figure Captions

Figure 1. A diffusion model of a two-alternative forced choice task.

Figure 2. Two realizations of the 2DSD model of confidence.

Figure 3. Possible distributions of covert confidence with the mapping hypothesis of the Poisson model.

Figure 4. Latency-confidence-choice probability functions Participants 3 and 4 during the line length discrimination task.

Figure 5. The Latency-confidence-choice probability functions Participants 3 and 4 during the city population discrimination task.

Figure 6. 2DSD model with trial variability fits to latency-confidence-choice probability functions for the line length discrimination task for all participants.

Figure 7 2DSD model with trial variability fits to latency-confidence-choice probability functions for the city population discrimination task for each participant.

Figure 8. Contour plot of the best fitting joint distribution of observed decision time t_D' by evidence at the time of the confidence rating $L(t_C)$ in the fourth level of difficulty (32 vs 33.87 mm).

Figure 9. Empirical and best fitting (model) calibration curves for the average participant in the line length (top row) and city population (bottom row) discrimination task.

Figure 10. Predicted Brier Scores for Participant 6 in City Population Task in Difficulty Level 4.

Figure 11. Observed and best fitting inter-judgment times (τ) as a function of confidence level in the line length task.

Figure 12. A Markov-chain approximation of a more general process 2DSD model of confidence ratings.

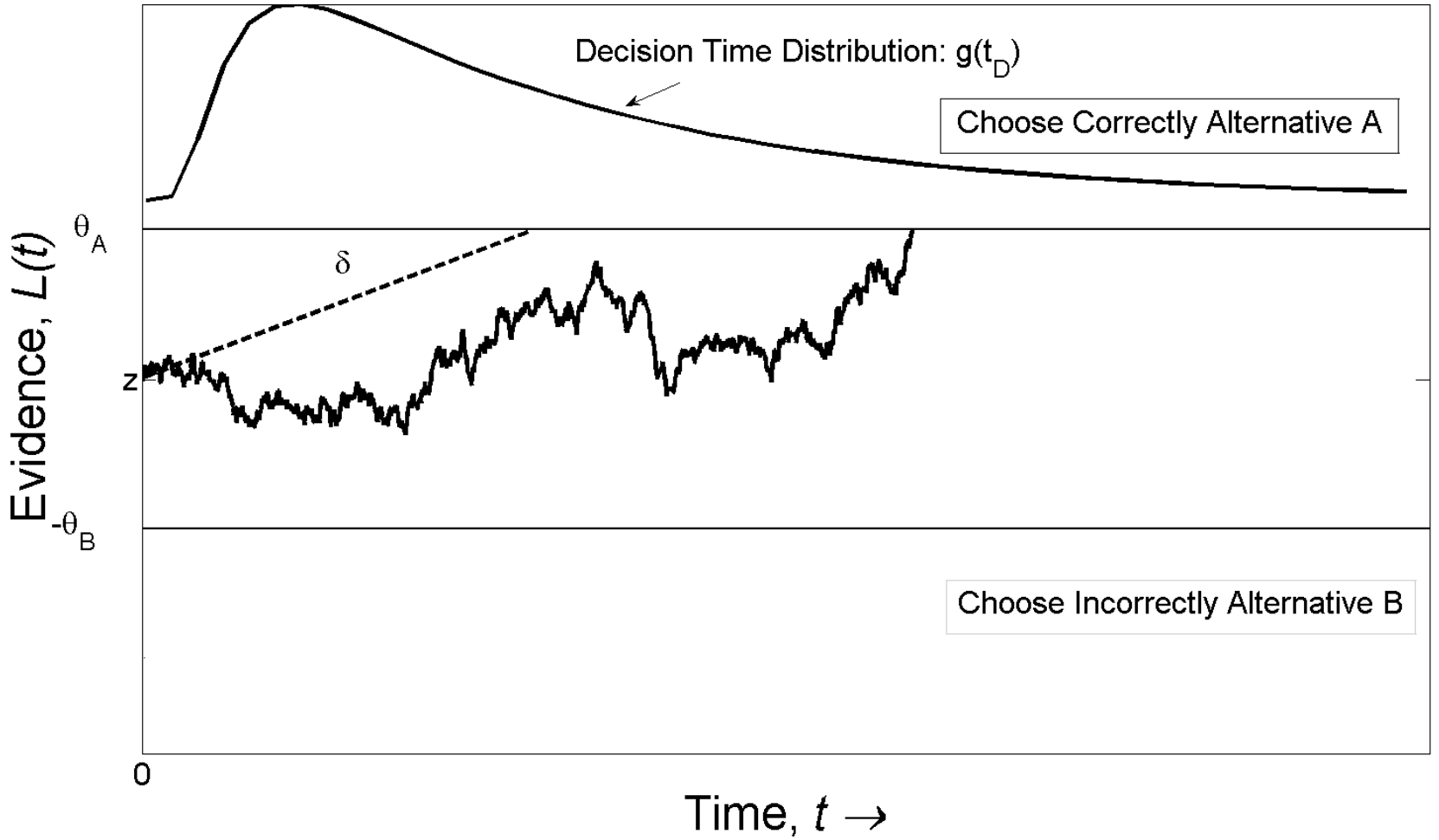


Figure 1. A diffusion model of a two-alternative forced choice task. The jagged line depicts a realization of a diffusion process, which represents the accumulated evidence on a given trial. The distribution at the top of the figure, $g(t_D)$, illustrates the predicted decision time distribution (first passage) when A is chosen for the diffusion model with these parameters.

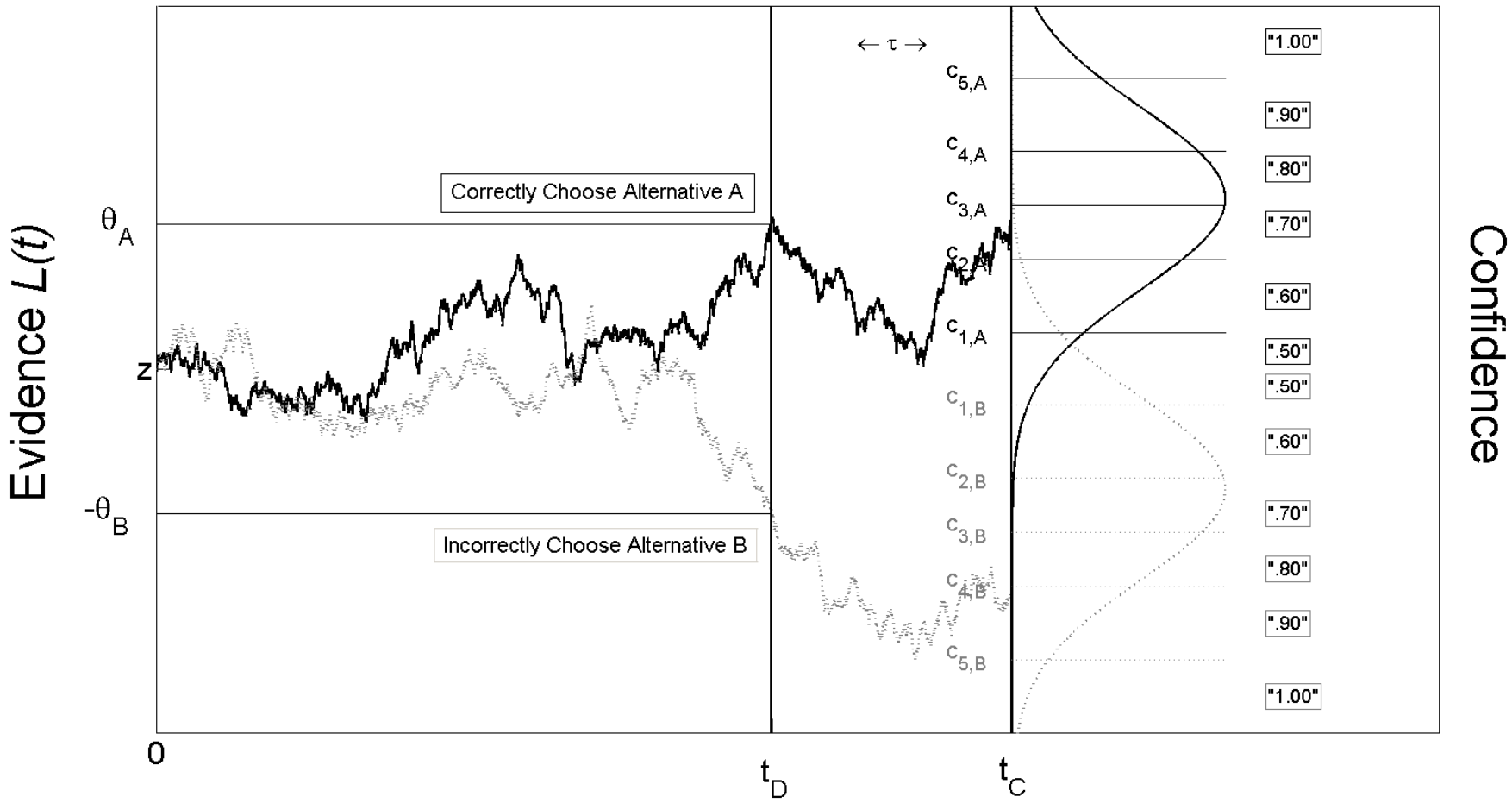


Figure 2. Two realizations of the 2DSD model of confidence. The black lines depict the process and the predicted distribution of confidence ratings when a judge correctly predicts choice alternative A. The light gray dotted lines depict the model when a judge incorrectly chooses alternative B. To produce a confidence estimate the model assumes after a fixed time interval passes or the inter-judgment time τ 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.

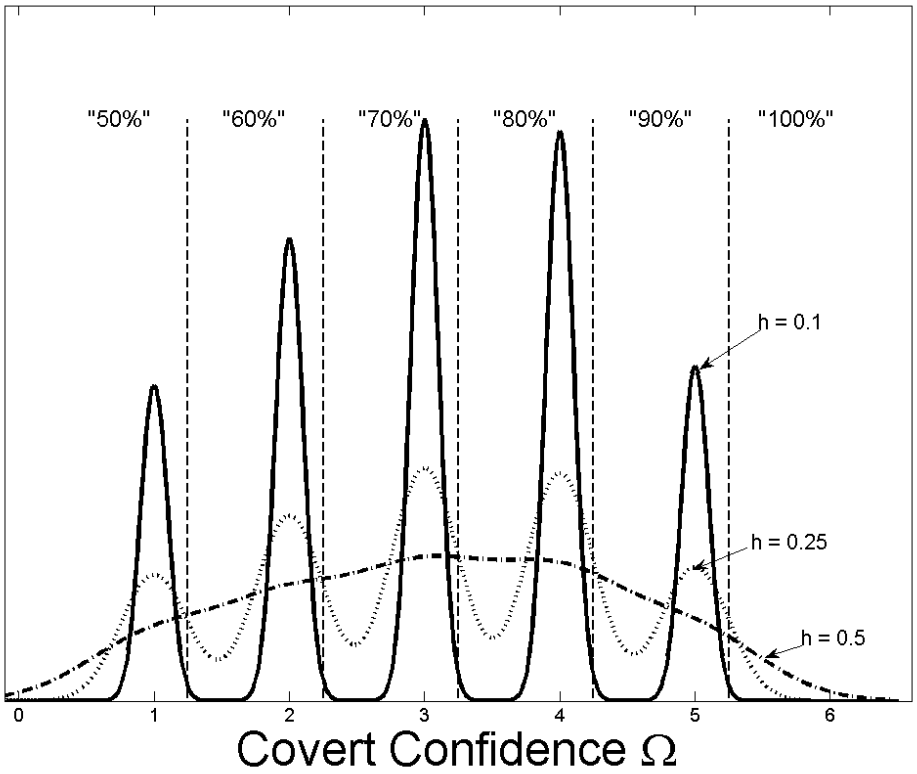


Figure 3. The distribution of covert confidence with the mapping hypothesis of the Poisson model. In the model, the balance of evidence counts, C, are a sample of possible values representing a larger population of graded levels of covert confidence, Ω. To estimate the distribution over Ω we assumed that there is a Gaussian density centered over C so that height of the density is a weighted average of all the observations in the sample. The parameter h is called a bandwidth parameter, which essentially controls the width of the Gaussian distribution placed over each sampled observation, C. The larger h gets the more normal the distribution over Ω becomes and the less the distribution reflects the discrete counts of the balance of evidence hypothesis in the Poisson model.

Figure 4. The Latency-confidence-choice probability functions Participants 3 and 4 during the line length discrimination task. The best fitting functions for the 2DSD and Poisson race model balance of evidence mapping hypothesis are shown. The circles with solid lines are the data, squares with dashed lines are the fits of the 2DSD model, and the triangles with dotted lines are the fits of the Poisson race model using the balance of evidence mapping hypothesis. Unshaded markers are the error or incorrect responses. Shaded markers are the correct responses. The error bars represent 95% confidence intervals.

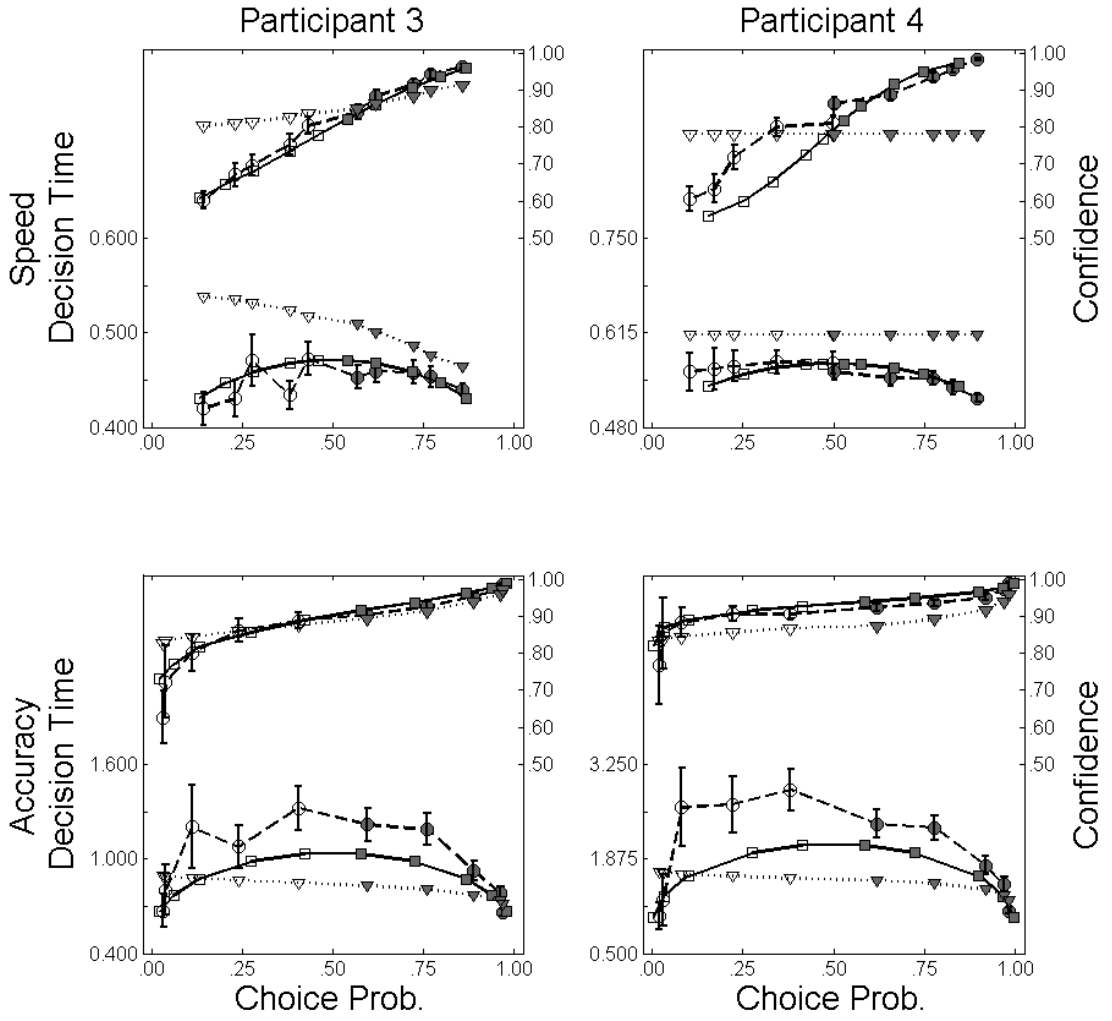


Figure 5. The Latency-confidence-choice probability functions Participants 3 and 4 during the city population discrimination task. The best fitting functions for the 2DSD and Poisson race model balance of evidence mapping hypothesis are shown. The circles with solid lines are the data, squares with dashed lines are the fits of the 2DSD model, and the triangles with dotted lines are the fits of the Poisson race model using the balance of evidence mapping hypothesis. Unshaded markers are the error or incorrect responses. Shaded markers are the correct responses. The error bars represent 95% confidence intervals.

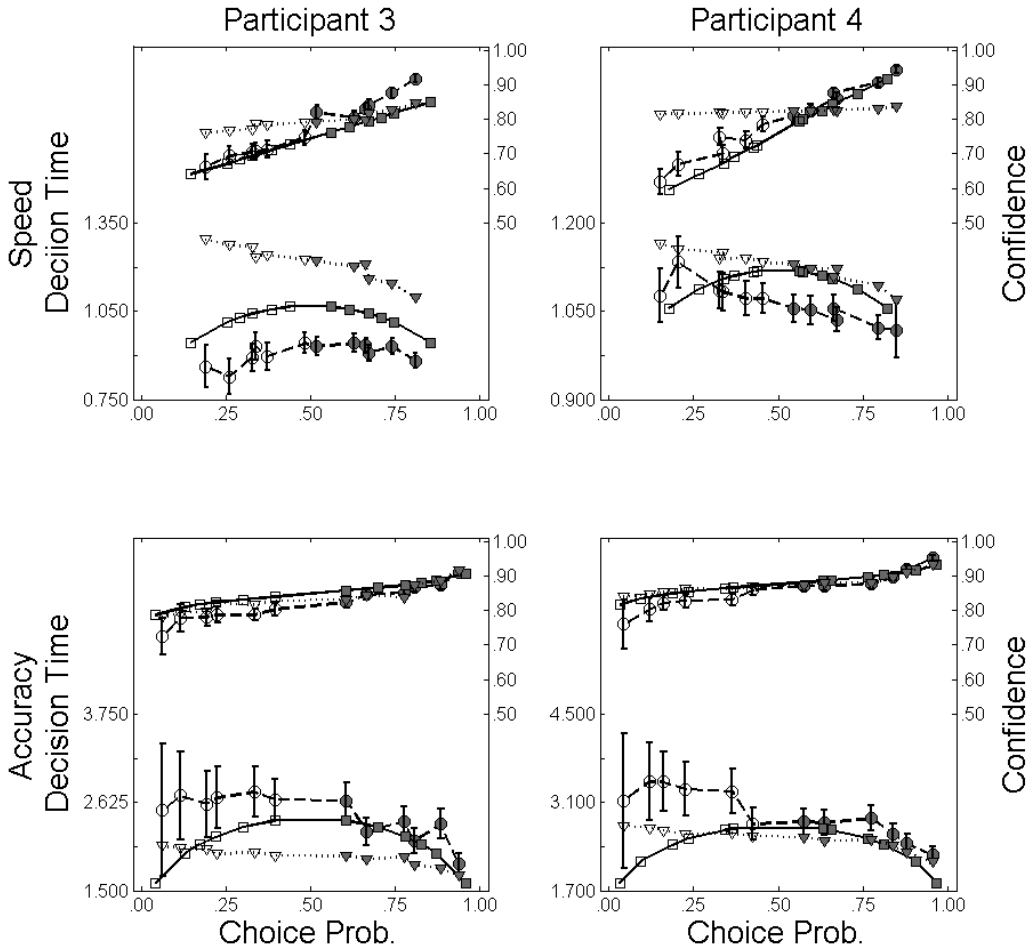


Figure 6. 2DSD model with trial variability fits to latency-confidence-choice probability functions for the line length discrimination task for all participants. The circles with solid lines are the data, squares with dashed lines are the fits of 2DSD model with trial variability in the parameters. The error bars represent 95% confidence intervals.

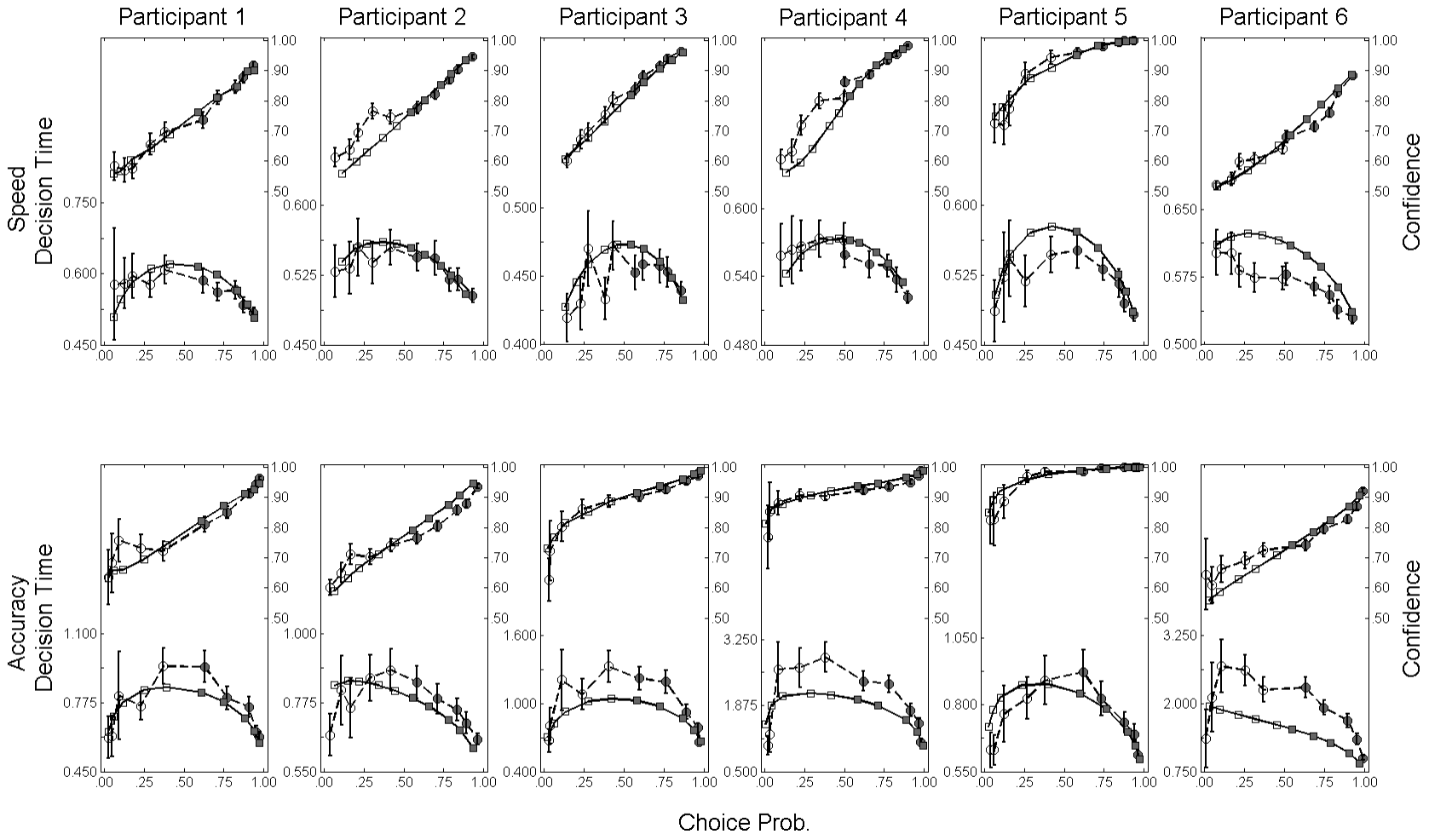


Figure 7 2DSD model with trial variability fits to latency-confidence-choice probability functions for the city population discrimination task for each participant. The circles with solid lines are the data, squares with dashed lines are the fits of 2DSD model with trial variability in the parameters. The error bars represent 95% confidence intervals.

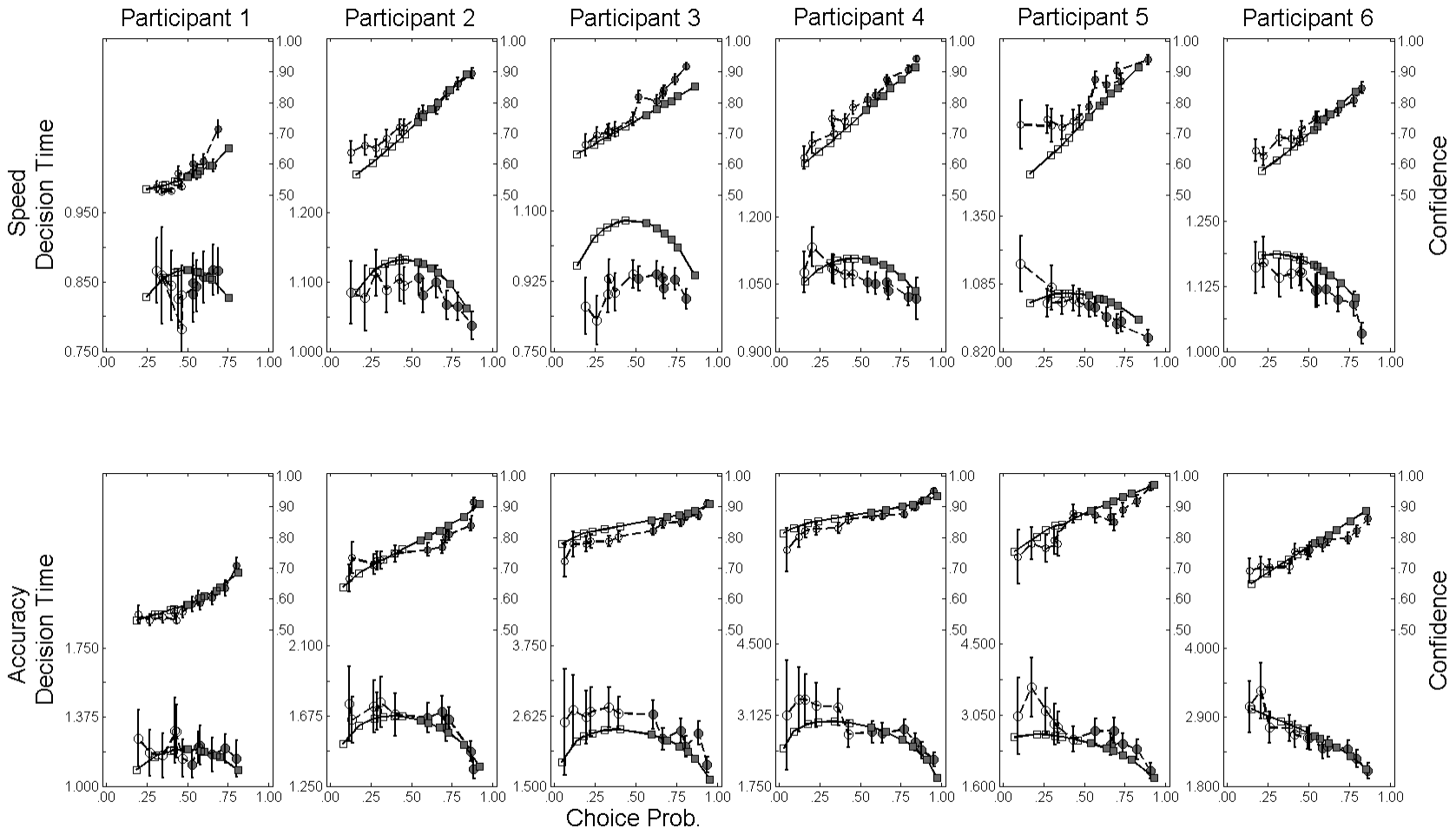


Figure 8. Contour plot of the best fitting joint distribution of observed decision time t_D' by evidence at the time of the confidence rating $L(t_C)$ in the fourth level of difficulty (32 vs 33.87 mm). The observed Goodman and Kruskal γ was -.32 (SE = .04) while the best fitting 2DSD estimated γ was -.24. The dotted lines display the confidence criteria. Below each contour plot are the observed and fitted (model) marginal distributions of decision times. The empirical distribution of decision times was inferred using a Gaussian kernel estimate (Van Zandt, 2000a). To the right are the observed and fitted (model) marginal distributions of confidence ratings.

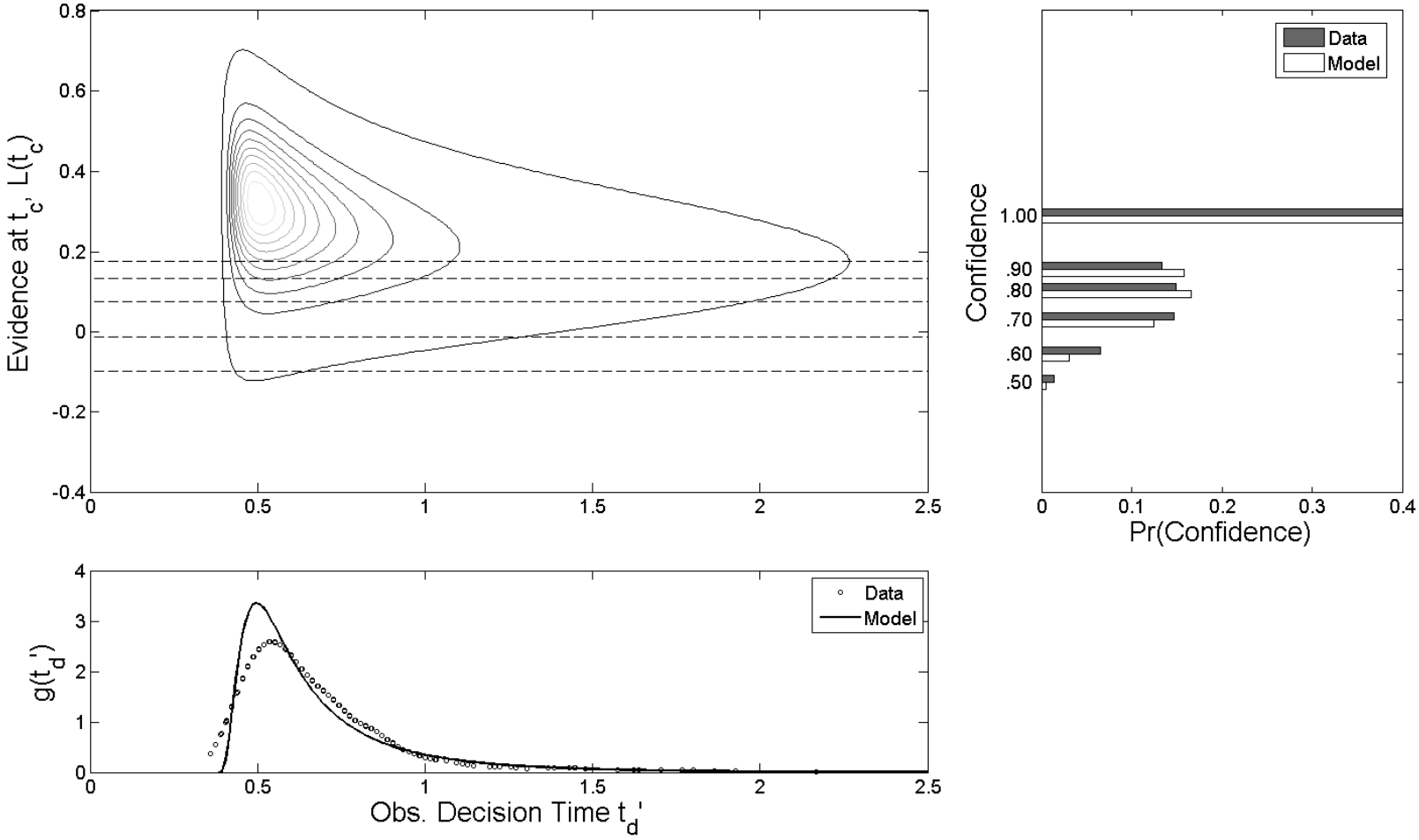


Figure 9. Empirical and best fitting (model) calibration curves for the average participant in the line length (top row) and city population (bottom row) discrimination task. The easy condition is the easiest 3 levels and the hard condition is the hardest 3 levels in the respective tasks. The error bars represent 95% confidence intervals calculated using the standard error of the proportion correct conditional on the confidence rating category.

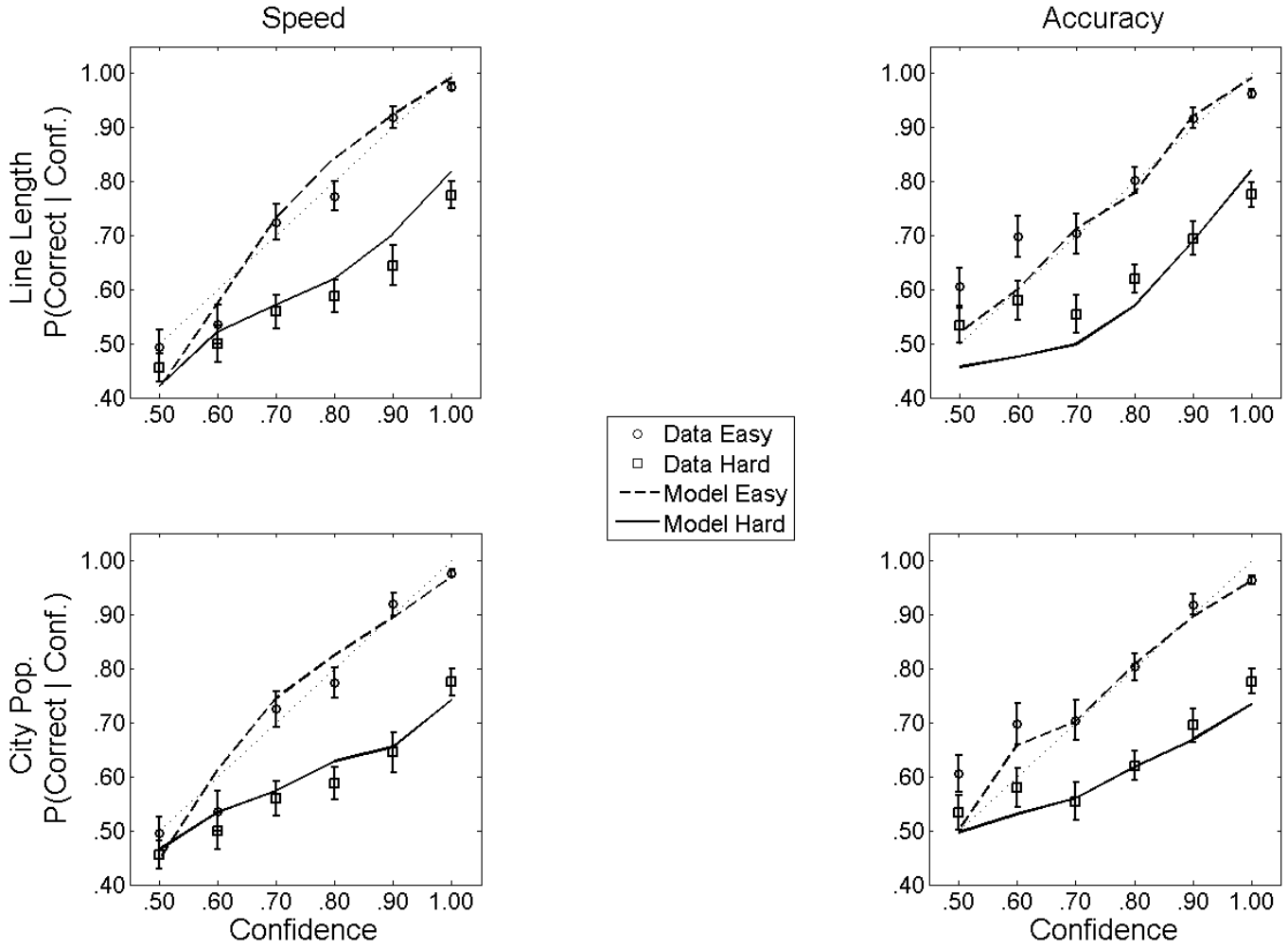


Figure 10. Predicted Brier Scores for Participant 6 in City Population Task in Difficulty Level 4. The plot was calculated using the 2DSD model without trial variability in the parameters. The plot illustrates that according to the 2DSD model increases in the choice threshold θ and inter-judgment time τ 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.

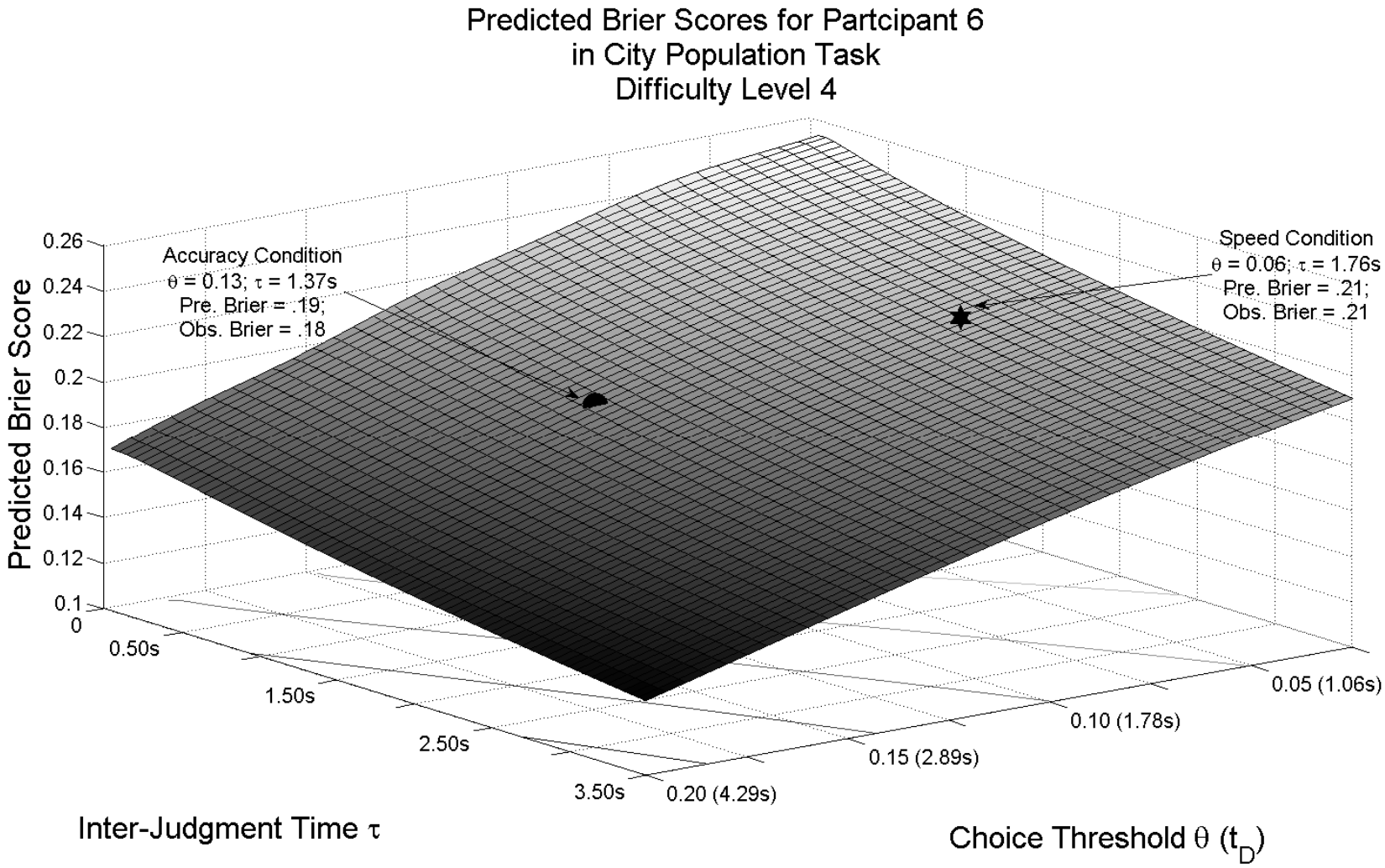


Figure 11. Observed and best fitting (model) inter-judgment times (τ) as a function of confidence level in the line length task. Inter-judgment times for both correct and incorrect on average grew faster with increasing levels of confidence. The figure shows that if the 2DSD model is formulated as a Markov chain and treating the confidence rating as an optional stopping response, then the model can account for this pattern of generally decreasing inter-judgment times with increasing levels of confidence.

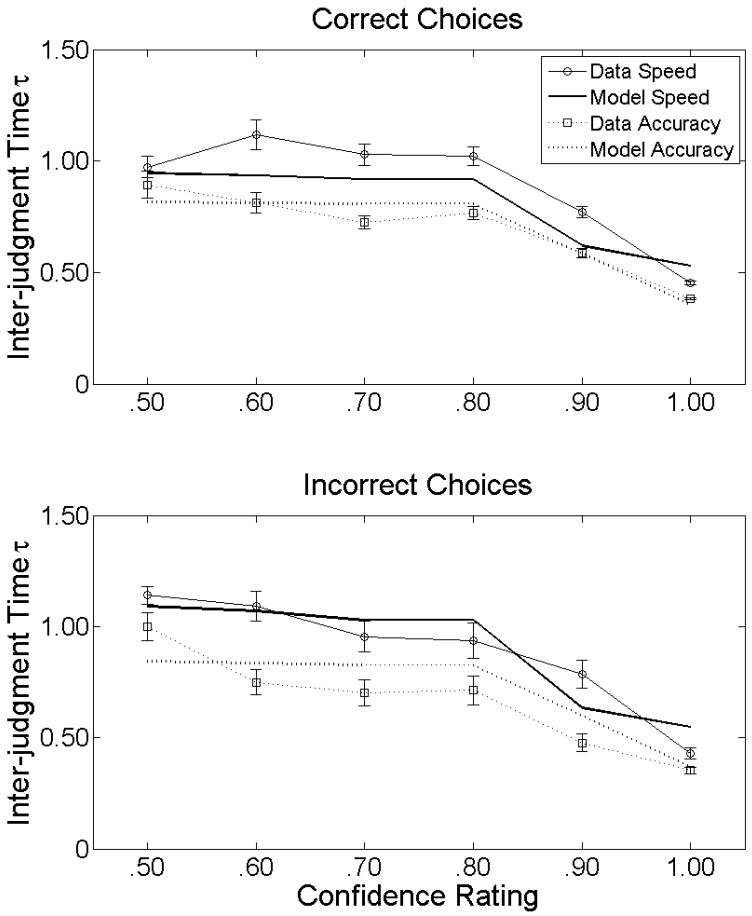
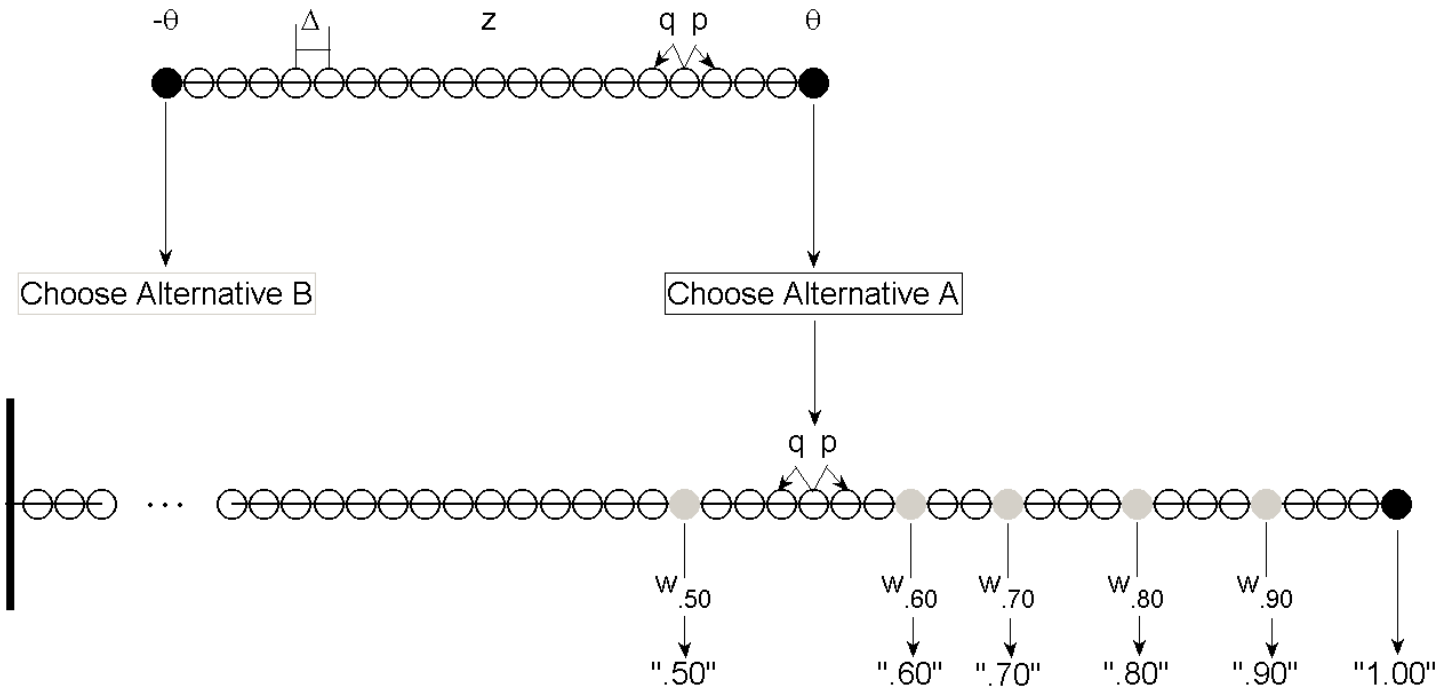


Figure 12. A Markov-chain approximation of a more general process 2DSD model of confidence ratings.



In the model, evidence accumulates over time toward an upper, θ, and lower threshold, - θ. This accumulation is approximated with discrete states in the model using probabilities p and q of moving a step size Δ to each adjacent state. This process can produce a trajectory such as the jagged line in Figure 1 or 2, producing a drift rate of δ toward either threshold. After making a choice, judges continue accumulating evidence, but are assumed to lay out markers across the state space so that if the process crosses through that particular state judges exit with probability w and give the corresponding confidence rating. To adequately fit the data, the model assumed that the confidence level of 1.00 was associated with an absorbing boundary so that if the process entered its associated state the judge stops accumulating evidence and states a "1.00" level of confidence. A reflecting boundary (straight line at far left) was placed at the other end of the chain so that if the evidence accumulation process hit this state it was reflected back. If alternative B was chosen a similar chain is used (not shown). This chain 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, using a Markov chain approximations of the diffusion model (see Appendix D).