Psychological Review
2010, Vol. 117, No. 3, 864–901

© 2010 American Psychological Association
0033-295X/10/$12.00 DOI: 10.1037/a0019737

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

Timothy J. Pleskac
Michigan State University

Jerome R. Busemeyer
Indiana University

The 3 most often-used performance measures in the cognitive and decision sciences are choice, response or decision time, and confidence. We develop a random walk/diffusion theory—2-stage dynamic signal detection (2DSD) theory—that accounts for all 3 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 3 indices of performance. Furthermore, the model also accounts for the distributions of each variable in both a perceptual and general knowledge task. The dynamic nature of the model also reveals the moderating effects of time pressure on the accuracy of choice and confidence. Finally, the model specifies the optimal solution for giving the fastest choice and confidence rating for a given level of choice and confidence accuracy. Judges are found to act in a manner consistent with the optimal solution when making confidence judgments.

Keywords: confidence, diffusion model, subjective probability, optimal solution, time pressure

Confidence has long been a measure of cognitive performance used to chart the inner workings of the mind. For example, in psychophysics confidence was originally thought to be a window onto Fechner's perceived interval of uncertainty (Pierce, 1877). At the higher levels of cognition, confidence ratings about recognition are used to test and compare different theories of memory (e.g., Ratcliff, Gronlund, & Sheu, 1992; Squire, Wixted, & Clark, 2007; Yonelinas, 1994). Confidence has also been used in the decision sciences to map the correspondence between people'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 responses on a test (Lichtenstein, Fischhoff, & Phillips, 1982), or the accuracy of local sports fans predicting the outcome of games (Yates & Curley, 1985).

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

Such a dynamic understanding of confidence has a number of applications. In this article, we use our dynamic understanding of confidence to address an important question involving confidence: What is the effect of time and time pressure on the accuracy of our confidence ratings? To address this question we use methods developed in the decision sciences where confidence ratings are treated as subjective probabilities (Adams,

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

A National Institute of Mental Health Research Service Award (MH019879) awarded to Indiana University supported both the beginning and finishing of this work. Jerome R. Busemeyer was supported by the National Science Foundation under Grant 0817965. Various components of this article were presented at the 2007 Annual Meeting for the Cognitive Science Society; the 2007 Annual Meeting for the Society for Mathematical Psychology; the 2008 Annual Meeting for the Society for Mathematical Psychology; the 2008 Annual Meeting for the Society for Judgment and Decision Making; the 2009 Biennial Conference on Subjective Probability, Utility, and Decision Making; and the 2009 Annual Conference of the Psychonomic Society. We thank Roger Ratcliff, Jim Townsend, Trish Van Zandt, Thomas Wallsten, and Avi Wershbale for their input on this work. We also thank members of Timothy J. Pleskac's 2010 Nature and Practice of Cognitive Science class for comments on the manuscript. We are also appreciative of Kate LaLonde and Kayleigh Vandenbussche for their assistance in data collection.

Correspondence concerning this article should be addressed to Timothy J. Pleskac, Department of Psychology, Michigan State University, East Lansing, MI 48823. E-mail: tim.pleskac@gmail.com

1957; Adams & Adams, 1961; Lichtenstein et al., 1982; Tversky & Kahneman, 1974). The accuracy of subjective probabilities has been well studied in the decision sciences (for reviews see Arkes, 2001; Griffin & Brenner, 2004; Koehler, Brenner, & Griffin, 2002; McClelland & Bolger, 1994). Yet, little is known as to how or why the accuracy of subjective probability estimates might change under time pressure and more generally how judges balance time and accuracy in producing not only choice but also confidence ratings.

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

In the first stage (left of the vertical line located at t_D in Figure 1), judges make a choice based on a sequential sampling process best described by random walk/diffusion theory (Laming, 1968; Link & Heath, 1975; Ratcliff, 1978; Stone, 1960). During this stage, judges begin to sequentially accumulate evidence favoring one alternative over the other. Typically, the evidence has some direction or *drift*. The path of the accumulated evidence is shown as a jagged line in Figure 1. If the evidence drifts upward it favors Response Alternative A and if it drifts downward it favors B. The information serving as evidence can come from one of many different sources including current sensory inputs and/or memory stores. The jagged line in Figure 1 also illustrates that the sampled evidence at each time step is subject to random fluctuations. This assumption also characterizes the difference between random walk and diffusion models. In random walk models the evidence is sampled in discrete time intervals, whereas in diffusion models the evidence is sampled continuously in time.

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

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

In summary, random walk/diffusion theory describes both choice and decision times as a compromise between the (a) the *quality* of the accumulated evidence as indexed by the drift rate of the evidence and (b) the *quantity* of the accumulated evidence as indexed by the choice thresholds (Ratcliff & Smith, 2004). Models

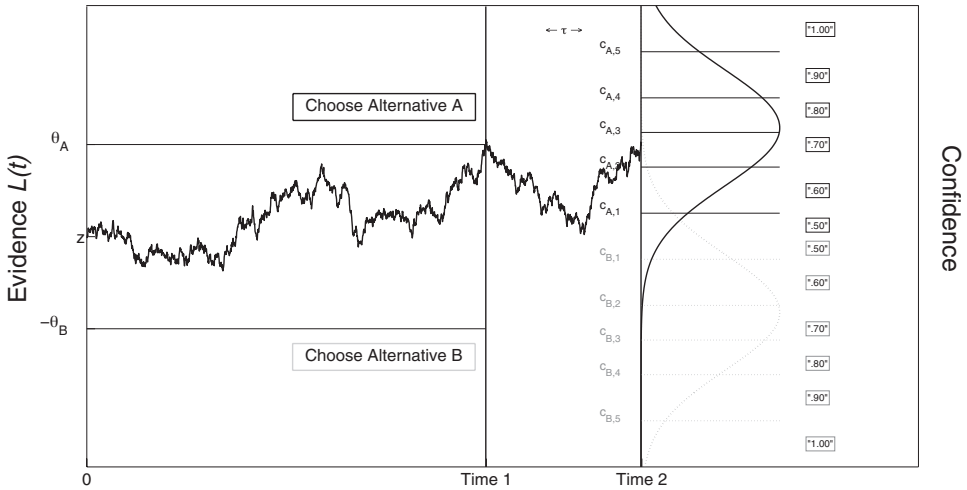

Figure 1. A realization of evidence accumulation in the two-stage dynamic signal detection (2DSD) interrogation model of confidence. The black jagged line depicts the accumulation process when a judge correctly predicts Response Alternative A. To produce a confidence estimate the model assumes after a fixed time interval passes (or interjudgment 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. The solid black normal curve on the right-hand side of the figure is the distribution of evidence at the time of confidence t_C when a judge correctly chooses Alternative A. The dashed normal curve is what the distribution would be if a judge would have incorrectly chosen Alternative B. θ_A and $-\theta_B$ = choice thresholds for Alternatives A and B, respectively; t_D = predicted decision time; z = starting point; $c_{choice,k}$ = confidence criteria.

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), multiattribute, multialternative decisions (Diederich, 1997; Roe, Busemeyer, & Townsend, 2001), as well as other types of response tasks like the go/no-go task (Gomez, Perea, & Ratcliff, 2007).

Often, though, an alternative measure of cognitive performance in the form of confidence is collected. Confidence is easily obtained with a simple adjustment in the empirical procedure: After judges make a choice, ask them to rate their confidence that their choice was correct. Indeed this simple empirical adjustment to collect confidence ratings has allowed psychologists to gain insight into a variety of areas including estimating empirical thresholds in psychophysics (Link, 1992; Pierce & Jastrow, 1884) and examining the role of consciousness and our own awareness of our cognitions (Nelson, 1996, 1997; Nelson & Narens, 1990), as well as simply being used as a method to collect data in an efficient manner for a variety of tasks (Egan, 1958; Green & Swets, 1966; Ratcliff et al., 1992).

In this article, we focus on the confidence with choice task. This focus allows us to explicitly model how choice and decision time are related to confidence. Moreover, by simultaneously modeling choice, decision time, and confidence, we investigate the degree to which the same dynamic process can account for all three measures of cognitive performance. Of course confidence could be collected in a slightly different manner, with a no-choice task where experimenters simply ask judges to rate their confidence that a particular item (e.g., A) is correct. In the discussion, we address how our model might be adapted to account for these tasks as well (see also Ratcliff & Starns, 2009).

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

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

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

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

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

Dynamic Signal Detection Theory

A common model of decision making is Green and Swets's (1966) signal detection theory. Random walk/diffusion theory can in fact be understood as a logical extension of signal detection theory (Busemeyer & Diederich, 2010; Link & Heath, 1975; Pike, 1973; Ratcliff & Rouder, 2000; Smith, 2000; Wagenmakers, van der Maas, & Grasman, 2007). In particular, the decision process in signal detection theory can be understood as using a fixed sample size of evidence to elicit a decision (akin to the stopping rule for interrogation choice tasks). Random walk/diffusion theory, in comparison, drops this assumption. Because of this logical connection to signal detection theory we have adopted the name *dynamic signal detection theory* to describe this more general framework of models. Formally, dynamic signal detection (DSD) theory assumes that as each time interval Δt passes after stimulus S_i (i = A, B) is presented, judges consider a piece of information $y(t)$. After a time length of $t = n(\Delta t)$ judges will have generated a

Table 1
Eight Empirical Hurdles a Model of Cognitive Performance Must Explain

Hurdle	Description	References
1. Speed–accuracy trade-off	Decision time and error rate are negatively 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 discriminability	Confidence increases monotonically as stimulus discriminability increases.	Ascher (1974); 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 positively 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	There is a monotonically increasing relationship between confidence and decision time where participants are on average more confident in conditions when they take more time to make a choice. This relationship is seen when comparing confidence across different conditions manipulating decision time (e.g., different stopping points in an interrogation paradigm or between speed and accuracy conditions in optional stopping tasks).	Irwin et al. (1956); Vickers & Packer (1982); Vickers, Smith, et al. (1985)
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); Swensson & 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 article; Baranski & Petrusic (1994)

set of n pieces of information drawn from some distribution $f_i[y(t)]$ characterizing the stimulus. Judges are assumed to transform each piece of information into evidence favoring one alternative over the other, $x(t) = h[y(t)]$. Because $y(t)$ is independent and identically distributed, $x(t)$ is also independent and identically distributed. Each new sampled piece of evidence $x(t + \Delta t)$ updates the total state of evidence $L(t)$ so that at time $t + \Delta t$ the new total state of the evidence is

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

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

$$dL(t) = L(t + \Delta t) - L(t) = x(t + \Delta t) = \delta\Delta t + \sqrt{\Delta t} \cdot \varepsilon(t + \Delta t), \quad (2)$$

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

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

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

$$E[L(t)] = \mu(t) = n \cdot \Delta t \cdot \delta = t \cdot \delta, \quad (3)$$

and so does the variance,

$$\text{var}[L(t)] = \sigma^2(t) = n \cdot \Delta t \cdot \sigma^2 = t \cdot \sigma^2, \quad (4)$$

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

$$d'(t) = 2\mu(t)/\sigma(t) = 2(\delta/\sigma)\sqrt{t} = d\sqrt{t}. \quad (5)$$

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

To make a choice, evidence is accumulated to either the upper (θ_A) or lower ($-\theta_B$) thresholds. Alternative A is chosen once the accumulated evidence crosses its respective threshold, $L(t) > \theta_A$. Alternative B is chosen when the process exceeds the lower threshold, $L(t) < -\theta_B$. The time it takes for the evidence to reach either threshold or the first passage time is the predicted *decision time*, t_D. This first passage account of decision times explains the positive skew of response time distributions (cf. Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006; Ratcliff & Smith, 2004). The model accounts for biases judges might have toward a choice alternative with a parameter z, the state of evidence at time point 0, $z = L(0)$. In this framework, if $z = 0$ observers are unbiased, if $z < 0$ then observers are biased to choose Alternative B, and if $z > 0$ then they are biased to respond hypothesis alternative A.[3]

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

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

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

Sequential Probability Ratio Tests (SPRT)

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

$$x(t) = h[y(t)] = \ln\left[\frac{f_A[y(t)]}{f_B[y(t)]}\right]. \quad (6)$$

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

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

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

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

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

[2] This unbridled growth of accuracy is also seen as an unrealistic aspect of random walk/diffusion models. More complex models such as Ratcliff's (1978) diffusion model with trial-by-trial variability in the drift rate and models with decay in the growth of accumulation of evidence (Ornstein Uhlenbeck models; Bogacz et al., 2006; Busemeyer & Townsend, 1993; Usher & McClelland, 2001) do not have this property.

[3] Ratcliff's (1978; Ratcliff & Smith, 2004) diffusion model places the lower threshold $-\theta_B$ at the zero point and places an unbiased starting point at the halfway point between the upper and lower thresholds.

The SPRT diffusion model has some empirical validity. Stone (1960) and Edwards (1965) used the SPRT diffusion model to describe human choice and decision times, although they failed to explain differences between mean correct and incorrect decision times (Link & Heath, 1975; Vickers, 1979). Gold and Shadlen (2001, 2002) have also worked to connect the SPRT model to decision making at the level of neuronal firing.

In terms of confidence, the model naturally predicts confidence if we assume judges transform their final internal posterior log odds (see Equation 8) with a logistic transform to a subjective probability of being correct. However, this rule (or any related monotonic transformation of the final log odds into confidence) implies that confidence is completely determined by the threshold values (θ_A or $-\theta_B$) or the quantity of evidence needed to make a choice. This predicted relationship between confidence and choice thresholds is problematic because when choice thresholds remain fixed across trials then this would imply that "all judgments (for a particular choice alternative) should be made with an equal degree of confidence" (Vickers, 1979, p. 175).

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

A further difficulty for the SPRT account of confidence is that the resolution of confidence is usually good. That is, judges' confidence ratings discriminate between correct and incorrect responses (e.g., Ariely et al., 2000; Baranski & Petrusic, 1998; Dougherty, 2001; Garrett, 1922; Henmon, 1911; D. M. Johnson, 1939; Nelson & Narens, 1990; Vickers, 1979). This resolution remains even when stimulus difficulty is held constant (Baranski & Petrusic, 1998; Henmon, 1911). In particular, judges typically have greater confidence in correct choices than in incorrect choices (Hurdle 3). The SPRT model, however, predicts equal confidence for correct and incorrect choices when there is no response bias.[4]

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

Pierce's Model of Confidence

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

$$\overline{conf} = \beta \cdot \ln\left[\frac{P(R_A|S_A)}{P(R_B|S_A)}\right]. \quad (9)$$

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

$$\overline{conf} = \beta \cdot \ln\left[\frac{P(R_A|S_A)}{P(R_B|S_A)}\right] \Big/ 2 = \delta\theta/\sigma^2. \quad (10)$$

In words, Pierce's hypothesis implies confidence is a multiplicative 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 DSD (for a more general derivation allowing for response bias, see Heath, 1984). For the remainder of this article, this function in combination with a DSD model describing choice and decision time is called *Pierce's model*. Link (2003) and Heath (1984) showed that Pierce's model gave a good account of mean confidence ratings.

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

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

Despite this setback, Pierce's model does bring to light two additional hurdles that 2DSD and any model of confidence must surmount. Pierce's model predicts that there is a negative relationship between decision time and the degree of confidence expressed

[4] The SPRT model could account for a difference in average confidence in correct 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 stimuli (i.e., hits and false alarms), though we are unaware of any empirical study directly testing this specific prediction.

in the choice. This is because as drift rate decreases the average decision time increases, while according to Equation 10 confidence decreases. This empirical prediction has been confirmed many times where across trials under the same conditions the average decision time monotonically decreases as the confidence level increases (e.g., Baranski & Petrusic, 1998; Festinger, 1943; D. M. Johnson, 1939; Vickers & Packer, 1982). This negative relationship between decision time and confidence in optional stopping tasks serves as empirical Hurdle 4 in Table 1.

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

Empirically a positive relationship between decision times and confidence was first identified in interrogation choice paradigms where an external event interrupts judges at different sample sizes and asks them for a choice. Irwin, Smith, and Mayfield (1956) used an expanded judgment task to manipulate the amount of evidence collected before making a choice and confidence judgment. An expanded judgment task externalizes the sequential sampling process by asking people to physically sample observations from a distribution and then make a choice.[5] As judges were required to take more observations, hence greater choice thresholds and longer decision times, their confidence in their choices increased (for a replication of the effect see Vickers, Smith, Burt, & Brown, 1985).

At the same time, when sampling is internal if we compare confidence between different levels of time pressure during optional stopping tasks, then we find that confidence is on average greater when accuracy as opposed to speed is a goal (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 model, which we return to shortly. Regardless, across these different tasks, this positive relationship between decision time and confidence eliminates many models of confidence and serves as Hurdle 5 for any model of confidence.

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

The Two-Stage Dynamic Signal Detection Model (2DSD)

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

There are several observations that support the assumptions of 2DSD. Neurophysiological studies using single cell recording techniques with monkeys suggest that choice certainty or confidence is based on the firing rate of the same neurons that also determines choice (Kiani & Shadlen, 2009). That is, confidence is a function of the state of evidence accumulation, $L(t)$. There is also support for the second stage of evidence accumulation. Anecdotally, we have probably all had the feeling of making a choice and then almost instantaneously new information comes to mind that changes our confidence in that choice. Methodologically some psychologists even adjust their methods of collecting confidence ratings to account for this postdecision 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 prespecified 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 rather continue collecting evidence.

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

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

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

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

$$L(t_C) = L(t_D) + \sum_{i=1}^{n} x(t_D + i \cdot \Delta t). \quad (11)$$

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

The distributions over the confidence ratings are functions of the possible evidence accumulations at time point t_C. The distribution of possible evidence states at time point t_C in turn reflects 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 the evidence at t_C is only a function of the evidence accumulated during the confidence period of time τ. Thus, based on Equation 3 and assuming evidence is accumulated continuously over time ($\Delta t \to 0$), when stimulus S_A is present the distribution of evidence at time t_C is normally distributed with a mean of

$$E[L(t_C)|S_A] = \begin{cases} \tau\delta + \theta_A, & \text{if } R_A \text{ was chosen} \\ \tau\delta - \theta_B, & \text{if } R_B \text{ was chosen} \end{cases}. \quad (12)$$

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

$$\text{var}[L(t_C)] = \sigma^2\tau. \quad (13)$$

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

$$\Pr(conf_j|R_A,S_A) = P(c_{R_A,j} < L(t_C) < c_{R_A,j+1}|\delta,\ \sigma^2,\ \theta_A,\ \tau), \quad (14)$$

where $c_{R_A,0}$ is equal to $-\infty$ and $c_{R_A,8}$ is equal to ∞. Similar expressions can be formulated for the other choices. The precise values of $\Pr(conf_j|R_A,S_A)$ can be found using the standard normal cumulative distribution function. Table 2 lists the parameters of the 2DSD model. The total number of parameters depends in part on the number of confidence ratings.

How Does the Model Stack Up Against the Empirical Hurdles?

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

Similarly, decreases in the drift rate also produce longer Stage 1 decision times and lower levels of confidence because confidence increases with drift rate. Thus, the model predicts a negative relationship between confidence and decision time (Hurdle 4). The 2DSD model also predicts a positive relationship between confidence and decision times in both optional stopping and interrogation paradigms (Hurdle 5). In optional stopping tasks, again judges set a larger threshold θ during accuracy conditions than in speed conditions. As a result this will move the means of the confidence distributions out, producing higher average confidence ratings in accuracy conditions. During interrogation paradigms average confidence increases as judges are forced to accumulate more evidence (or take more time) on any given trial. Within the 2DSD model this implies that the expected state of evidence will be larger because it is a linear function of time (see Equation 3), and thus confidence will be greater when judges are forced to take more time to make a choice.

Finally, 2DSD also accounts for a number of other phenomena. One example of this is an initially puzzling result where comparisons of confidence between speed and accuracy conditions showed that there was no difference in average confidence ratings

Table 2
Parameters of the Two-Stage Dynamic Signal Detection Interrogation Model of Confidence

Parameter	Meaning	Description
δ	Drift rate	Controls the average rate of evidence accumulation across time and indexes the average strength or quality of the evidence judges are able to accumulate. In fitting the model to the data, the drift rate was made a random variable drawn from a normal distribution with mean ν and variance η^2.
σ^2	Drift coefficient	Responsible for the within-trial random fluctuations. It is unidentifiable within a particular condition. In fitting the model, σ is set to .1.
θ_A, θ_B	Choice threshold	Determines the quantity of evidence judges accumulate before selecting a choice. Controls the speed–accuracy trade-off. In fitting the model, we set $\theta = \theta_A = \theta_B$.
z	Starting point	Determines the point in the evidence space where judges begin accumulating evidence. In fitting the model to the data, the starting point was made a random variable drawn from a uniform distribution centered at $z = 0$ with a range s_z.
t_{ED}	Mean nondecision time	Accounts for the nondecision time during the task (e.g., motor time). Observed decision time is a function of the nondecision 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 is one less confidence criterion than confidence levels.
τ	Interjudgment time	Indexes the time between when a decision is made and a confidence rating is entered.
t_{EJ}	Mean nonjudgment time	Accounts for the nonjudgment time during the task. Observed interjudgment time is a function of the nonjudgment time and interjudgment time used in the model, $\tau' = t_{EJ} + \tau$.
Trial variability parameters		
ν	Mean drift rate across trials	Indexes the mean quality of evidence across trials assuming a normal distribution.
η	Standard deviation of drift rate across trials	Indexes the variability of the quality of evidence across trials assuming a normal distribution.
s_z	Range of starting points	The range of starting points for the uniform distribution. In fitting the model, this parameter was constrained to be no larger than the smallest choice threshold.

between the two conditions (Festinger, 1943; Garrett, 1922; D. M. Johnson, 1939). This result speaks to some degree against Hurdle 5. Vickers (1979) observed though 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 were manipulated between sessions prompted Vickers (1979) to hypothesize that participants spread their confidence ratings out across the scale within each session. As a result they used the confidence scale differently between sessions, and this in turn would lead to equal confidence across accuracy and speed conditions. In support of this prediction, Vickers and Packer (1982) found that when the "complete scale" instructions were used in tandem with manipulations of speed and accuracy within sessions, judges were less confident during speed conditions (though see Baranski & Petrusic, 1998). 2DSD naturally accounts for this result because it makes explicit—via confidence criteria—the process of mapping a confidence rating to the state of evidence at the time of the confidence rating.

An additional advantage of making the confidence mapping process explicit in 2DSD is that it does not restrict the model to a specific scale of confidence ratings. 2DSD can be applied to a wide range of scales long used in psychology to report levels of confidence in a choice, such as numerical Likert-type scales (1, 2, . . .), verbal probability scales (*guess*, . . . , *certain*), and 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 DSD model where after making a choice, judges continue to accumulate evidence in support of the two alternatives. They then use the complete set of evidence to estimate their confidence in their choice. We have shown that this basic model accounts for a wide range of historical findings (Hurdles 1–5 in Table 1). Accounting for these datasets is an important set of hurdles to clear because they have been used to rule out possible alternative theories of confidence rooted within random walk/diffusion theory as well as many other theories (Vickers, 1979, 2001). 2DSD also makes a number of new predictions. For example, as Figure 1 and Equations 12, 13, and 14 imply, 2DSD does generally predict that—all else being equal—increases in τ should increase both the mean difference between confidence ratings in correct responses and incorrect choices (*slope*) and the pooled variance of the distribution of confidence ratings across correct and incorrect choices (*scatter*). These are strong predictions, and at first glance intuition might suggest they are incorrect predictions. To test these predictions we used data collected in a new study where participants completed two often-studied but different two-alternative forced choice situations: perceptual and general knowledge.

Overview of Empirical Evaluation of 2DSD

During the study, six participants completed a perceptual task and a general knowledge task. In the perceptual task participants were shown one of six possible pairs of horizontal lines and asked to (a) identify which line was longer/shorter and then (b) rate their confidence in their response on a subjective probability scale (.50,

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

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

Method

Participants

The six 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 participation in each of the approximately 20 sessions. All the participants were right handed and had normal or corrected-to-normal vision. Participants earned between $10 and $14 for each session.

Apparatus

All stimuli were presented using software programmed in E-Prime 2.0 Professional. This allowed for controlled presentation of graphics, instructions, event sequencing, timing, and recording of responses. Participants recorded their responses using a standard Dell QWERTY keyboard with most of the keys removed save two rows. The first row contained the *V*, *B*, and *N* keys and were relabeled as ←, *H*, and →, 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, which were relabeled with confidence ratings *50*, . . . , *100* to correspond with the confidence ratings in percentage terms. Participants were instructed to use their dominant hand to enter all responses. Periodic inspections during each session confirmed that participants adhered strictly to this instruction. Participants sat in individual sound-attenuated booths approximately 60 cm away from the screen.

Response Entry Task

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

Line Length Discrimination Task

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

Design and procedure. Participants completed 10 consecutive experimental sessions for the line length discrimination task. During each session participants completed three tasks. The first task was the previously described response entry task. The second

task was a practice set of eight trials (a block of four accuracy trials and a block of four speed trials counterbalanced across sessions). The order of practice blocks was counterbalanced across participants and between sessions. The final task was the experimental task. During the experimental task, participants completed eight blocks of trials. During each block (speed or accuracy) participants completed 72 trials (6 line pairs × 2 presentation orders × 2 longer/shorter instructions × 3 replications). The longer/shorter instructions meant that for half the trials participants were instructed to identify the longer line and for the other half they were instructed to identify the shorter line. This set of instructions replicates conditions used by Baranski and Petrusic (1998), Henmon (1911), and others in the foundational studies of confidence. Half the participants began Session 1 with an accuracy block and half began with a speed block. Thereafter, participants alternated between beginning with a speed or an accuracy block from session to session. In total participants completed 2,880 accuracy trials and 2,880 speed trials.

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

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) the number of trials completed (out of 72) and the number of blocks completed (out of eight). When they were ready, participants pressed the *H* key, which (a) removed trial and block information, (b) moved the instruction to the top of the screen, and (c) put a fixation cross in the center. Participants were told to fixate on the cross and to press the *H* key when ready. This press removed the fixation cross and put the pair of lines in the center with the corresponding choice keys at the bottom (← or →). Once a choice was entered a confidence scale was placed below the corresponding keys and participants were instructed to enter their confidence that they chose the correct line (50%, 60%, . . . , 100%). After entering a confidence rating, feedback was given if they made an incorrect choice in the accuracy block or if they were too slow in the speed block. Otherwise no feedback was given and participants began the next trial.

At the beginning of the experiment, participants were told to select a confidence rating so that over the long run the proportion of correct choices for all trials assigned a given confidence rating should match the confidence rating given. Participants were reminded of this instruction before each session. This instruction is common in studies on the calibration of subjective probabilities (cf. Lichtenstein et al., 1982). As further motivation, participants earned points based on the accuracy of their choice and confidence rating according to the quadratic scoring rule (Stäel von Holstein, 1970),

$$\text{points} = 100[1 - (correct_i - conf_i)^2], \tag{15}$$

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 (.50, .60, . . . , 1.00). This scoring rule is a variant of the Brier score (Brier, 1950), and as such it is a strictly proper scoring rule ensuring that participants will maximize their earnings only if they maximize their accuracy in both their choice and their confidence rating. Participants were informed of the properties of this scoring rule prior to each session and shown a table demonstrating why it was in their best interest to accurately report their choice and confidence rating. To enforce time pressure during the speed conditions, the points earned were cut in half if a choice exceeded the deadline of 750 ms and then cut in half again every 500 ms after that. For every 10,000 points participants earned an additional $1.

City Population Discrimination Task

Stimuli. The city pairs were constructed from the 100 most populated U.S. cities according to the 2006 U.S. Census estimates. There are 4,950 pairs. From this population, 10 experimental lists of 400 city pairs were randomly constructed (without replacement). The remaining pairs were used for practice. During the choice trials, the city pairs were shown in the center of a black screen. The city names were written in yellow and centered around the word *or*, which was written in red. Immediately below each city was the state abbreviation (e.g., MI for Michigan).

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

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

Results

The Results section is divided into four sections. The first section summarizes the behavioral results from the two tasks and examines several qualitative predictions that 2DSD makes regarding the effect of changes in interjudgment time τ on confidence. These predictions are also of interest because race models using Vickers's (1979) balance of evidence hypothesis of confidence predict the opposite pattern of results. The second section examines the fit of the 2DSD interrogation model to the data. In this section, we also investigated the degree to which trial variability in the process parameters—a construct of interest to both cognitive (Ratcliff & Rouder, 1998) and decision scientists (Erev, Wallsten, & Budescu, 1994) alike—adds to the fit of the 2DSD model. In the third section, we provide an explanation for the trade-offs made between the entire time course of the judgment process (decision time + interjudgment time) and the accuracy of both the observed choice and confidence rating. This trade-off is important not only for decision scientists who have long focused on the accuracy of confidence judgments, but also because it can help explain the observed behavior of participants in our study. Finally, we present an evaluation of a more extensive version of 2DSD that offers a more precise process account of the distributions of interjudgment times.

As a first step in the data analyses, in both the line length and city population 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 s and the observed interjudgment times were less than 0.15 s. To minimize slow outliers, we excluded trials where either the decision time or observed interjudgment time was greater than 4 *SD*s from the mean. These cutoffs eliminated on average 2.5% (*min* = 1.1%; *max* = 4.9%) of the data in the line length task and 2.0% (*min* = 1.0%; *max* = 5.3%) of the data in the city population task.

Behavioral Tests of 2DSD

Between-conditions results. Table 3 lists the proportion correct, the average decision time, the average confidence rating, and the average interjudgment time for each participant in the line length and city population task. Throughout the article when statistics are listed averaged across participants they were calculated using methods from meta-analysis where each participant's data were treated as a separate experiment and the average statistic is calculated by weighting each participant's respective statistic by the inverse of the variance of the statistic (Shadish & Haddock, 1994). These estimates were calculated

Table 3
Proportion Correct, Average Decision Time, Average Confidence Rating, and Average Interjudgment Time for Each Participant

	Participant						
Task	1	2	3	4	5	6	*M*
Line length							
Proportion correct							
Speed	.82*	.80*	.73*	.76*	.83*	.78*	.79*
Accuracy	.87	.83	.86	.88	.85	.87	.86
Decision time							
Speed	0.54 (0.15)*	0.52 (0.10)*	0.45 (0.10)*	0.54 (0.09)*	0.51 (0.11)*	0.55 (0.09)*	0.52*
Accuracy	0.69 (0.33)	0.72 (0.30)	0.87 (0.57)	1.73 (1.47)	0.70 (0.33)	1.56 (1.24)	1.04
Confidence							
Speed	.82 (.22)*	.85 (.16)	.87 (.17)*	.90 (.15)*	.97 (.12)*	.76 (.18)*	.86*
Accuracy	.89 (.18)	.85 (.16)	.94 (.12)	.96 (.08)	.99 (.06)	.84 (.16)	.91
Interjudgment time							
Speed	0.52 (0.43)	0.61 (0.52)	0.58 (0.32)	1.05 (0.82)	0.31 (0.22)	1.02 (0.60)	0.68
Accuracy	0.50 (0.40)	0.52 (0.31)*	0.49 (0.27)*	0.40 (0.24)*	0.26 (0.11)*	0.69 (0.41)*	0.48*
City population							
Proportion correct							
Speed	.59*	.69*	.67*	.68*	.68*	.66	.66*
Accuracy	.64	.75	.78	.78	.73	.68	.73
Decision time							
Speed	0.84 (0.26)*	1.08 (0.17)*	0.91 (0.22)*	1.05 (0.17)*	0.96 (0.24)*	1.11 (0.19)*	0.99*
Accuracy	1.16 (0.56)	1.57 (0.58)	2.33 (1.52)	2.74 (1.59)	2.58 (1.74)	2.43 (1.16)	2.14
Confidence							
Speed	.58 (.16)*	.78 (.17)*	.81 (.14)*	.83 (.15)*	.83 (.20)*	.75 (.17)*	.76*
Accuracy	.60 (.17)	.80 (.17)	.85 (.12)	.89 (.10)	.87 (.17)	.78 (.16)	.80
Interjudgment time							
Speed	0.58 (0.48)	0.86 (0.42)	0.66 (0.46)	1.27 (0.78)	1.22 (0.80)	1.72 (0.82)	1.05
Accuracy	0.53 (0.39)*	0.57 (0.22)*	0.32 (0.17)*	0.34 (0.14)*	0.65 (0.44)*	1.35 (0.72)*	0.63*

Note. Decision time and interjudgment time were measured in seconds. Values in parentheses are standard deviations. An asterisk indicates the condition (speed or accuracy) in which a *z* test revealed the relevant statistic was smaller using an alpha value of .05 (two-tailed). The column at the far right lists the average value of the relevant statistic calculated by weighting each participant's respective statistic by the inverse of the variance of the individual statistic. Statistical significance for the average participant was determined using the average standard error, assuming random effects.

assuming a random effects model. Calculating the average statistic in this way provides an estimate of the standard error around the average statistic.

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

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

More broadly, though, we interpret this result of increased interjudgment time during the speed conditions as offering preliminary support for our hypothesis that judges continue to engage in postdecisional stimulus processing to enter their confidence rating. If no postdecisional processing occurred and instead choice and confidence are simultaneously available as most models assume, then one would expect no difference in interjudgment time between time pressure conditions. In the third section, we show that that this strategy of increasing interjudgment 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 Goodman and Kruskal's Γ ordinal measure of association (Goodman & Kruskal, 1954). Goodman and Kruskal's Γ assumes only an ordinal scale and makes no distribution assumptions (Goodman & Kruskal, 1954). 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 (Nelson, 1984, 1987).

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

Table 4
Average Goodman and Kruskal Γ Correlation Coefficients Across Each Participant for Both Tasks in the Speed and Accuracy Conditions

Task	Objective difference	Accuracy	Decision time	Confidence	Interjudgment time
Line length					
Objective difference between line lengths	—	.50 (.06)*	−.13 (.04)*	.34 (.10)*	−.15 (.07)*
Accuracy	.68 (.05)*	—	−.12 (.11)*	.75 (.09)*	−.31 (.10)*
Decision time	−.26 (.08)*	−.26 (.10)*	—	−.16 (.18)*	.14 (.06)*
Confidence	.40 (.16)*	.67 (.11)*	−.34 (.30)*	—	−.52 (.25)*
Interjudgment time	−.11 (.10)*	−.18 (.12)*	.22 (.11)*	−.47 (.25)*	—
City population					
Objective difference between city populations	—	.26 (.08)*	−.06 (.05)*	.18 (.04)*	−.07 (.05)*
Accuracy	.35 (.08)*	—	−.10 (.12)	.54 (.09)*	−.17 (.12)*
Decision time	−.10 (.04)*	−.17 (.08)*	—	−.14 (.16)*	.08 (.03)*
Confidence	.20 (.05)*	.43 (.06)*	−.35 (.15)*	—	−.26 (.27)*
Interjudgment time	−.03 (.05)	−.07 (.09)	.12 (.05)*	−.16 (.23)	—

Note. The values below the diagonal are the Γ coefficients for the accuracy condition, and the values above the diagonal are for the speed condition. The values in parentheses are an estimate of the between-participants standard deviation of the Γ coefficient. 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 assuming random effects.
* $p < .05$, two-tailed.

monotonically related to accuracy (Hurdle 3) and inversely related to decision time (Hurdle 4).[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 larger amount of variability from both the stimuli and the participant. Nevertheless, the pattern of associations is at least consistent with the hypothesis that although the information comes from a different source, a similar decision process is used in each task. The associations in Table 4 also reveal places where the 2DSD interrogation model is silent. In particular, in all conditions and all tasks interjudgment time and the confidence rating were negatively related (Γ = –.16 to –.52); we return to this result in the final results section.

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

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

$$slope = \overline{conf}_{correct} - \overline{conf}_{incorrect}. \quad (16)$$

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

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

Another relevant statistic is the scatter score (Yates, 1990) or the pooled variance of confidence across the correct and incorrect choices,

$$scatter = \frac{n_{correct}\mathrm{var}(conf_{correct}) + n_{incorrect}\mathrm{var}(conf_{incorrect})}{n_{correct} + n_{incorrect}}. \quad (17)$$

According to 2DSD, scatter should increase with longer interjudgment times (τ) because the variance of the distribution of evidence at the time of confidence increases with interjudgment time (see Equation 13). Table 5 lists the scatter score for each participant in the speed and accuracy conditions of both the line length and city population discrimination tasks. The scatter statistics show that for most participants (four out of six in the line length task and five out of six in the city population task) as well as the average participant, variance was significantly larger during the speed condition compared to the accuracy condition.[8] Note that unlike the slope score, changes in the choice threshold θ have little to no effect on the scatter of participants because θ influences only the mean confidence level, not the deviations from the mean.

[6] Pierce's (1877) model actually predicts a linear relationship between the proportion correct and the mean confidence rating (see Equation 9). In fact, the data from both tasks supported this prediction. In the line length task, regressing the average confidence rating across all six subjects onto the average proportion of correct choices accounts for 98% and 99% of the variance in the speed and accuracy conditions, respectively. To evaluate this prediction in the city population task, we first formed six groups of city pairs based on the difference in the ordinal ranking of the city pairs in terms of population (see model fitting section for more details). Using these groups, Pierce's model accounted for 95% and 90% of the variance in mean confidence rating in the speed and accuracy conditions, respectively.

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

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

Table 5
Scores of Slope and Scatter for Each Participant in the Speed and Accuracy Conditions of Each Task

	Participant						
Task	1	2	3	4	5	6	*M*
Line length							
Slope							
Speed	.22 (.01)	.17 (.01)	.21 (.01)	.20 (.01)	.13 (.01)	.21 (.01)	.19 (.01)
Accuracy	.19 (.01)*	.17 (.01)	.12 (.01)*	.07 (.00)*	.04 (.00)*	.16 (.01)*	.12 (.03)*
Scatter							
Speed	.039 (.001)	.022 (.001)	.019 (.001)	.015 (.001)	.011 (.001)	.025 (.001)	.022 (.003)
Accuracy	.028 (.001)*	.023 (.001)	.012 (.001)*	.005 (.000)*	.003 (.000)*	.024 (.001)	.016 (.005)*
City population							
Slope							
Speed	.08 (.01)	.13 (.01)	.14 (.01)	.15 (.01)	.14 (.01)	.10 (.01)	.12 (.01)
Accuracy	.08 (.01)	.10 (.01)*	.08 (.01)*	.06 (.01)*	.10 (.01)*	.08 (.01)*	.08 (.01)*
Scatter							
Speed	.023 (.001)*	.027 (.001)	.017 (.000)	.017 (.001)	.037 (.001)	.027 (.001)	.024 (.003)
Accuracy	.027 (.001)	.026 (.001)*	.012 (.000)*	.009 (.000)*	.028 (.001)*	.024 (.001)*	.021 (.005)*

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

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

Race models. Race models using Vickers's (1979) balance of evidence hypothesis actually predict a decrease in slope and scatter as time pressure increases. This class of models offers an alternative sequential sampling process to choice. The basic idea is that when judges are presented with a choice between two alternatives, evidence begins to accrue on counters, one for each response alternative. The first counter to reach a threshold determines the choice. Thus, choice in these models is based on absolute count of evidence as opposed to a relative amount of evidence as in DSD models (Ratcliff & Smith, 2004).

Models using this type of sequential sampling process include the linear ballistic accumulator model (Brown & Heathcote, 2008), the Poisson race model (Pike, 1973; Townsend & Ashby, 1983), the accumulator model (Vickers, 1979), and other biologically inspired models (Usher & McClelland, 2001; Wang, 2002). They appear to give a good account of choice data, though some models like the linear ballistic accumulator appear to do better at accounting for decision time distributions and changes in the distributions than others like the accumulator and Poisson race model (Brown & Heathcote, 2008; Ratcliff & Smith, 2004).[9]

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

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

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

A second stage of processing could be adopted in race models to account for these results. For instance, Van Zandt and Maldonado-Molina (2004) have developed a two-stage Poisson race model. However, we point out that the Poisson race model has several other difficulties it must overcome. For instance, the model has difficulty in accounting for distributions of decision times (Ratcliff & Smith, 2004).[10] There are other race models that might be better candidates for the two-stage balance of evidence hypothesis, including Brown and Heathcote's (2008) linear ballistic accumulator and Usher and McClelland's (2001) leaky accumulator model. We leave a full model comparison with these models for future investigation. Next we examine several new quantitative predictions 2DSD makes. To do so it is useful to fit the 2DSD model to the multivariate distribution of choice, decision time, and confidence.

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

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

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

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

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

These two hurdles—slow errors for difficult conditions (Hurdle 6) and fast errors for easy conditions (Hurdle 7)—are problematic for many sequential sampling models to explain. Ratcliff and colleagues have shown that in order to simultaneously account for this pattern of results DSD requires trial-by-trial variability in the process parameters (Ratcliff & Rouder, 1998; Ratcliff & Smith, 2004; Ratcliff et al., 1999). Trial-by-trial variability in processing stimuli is an often-made assumption in stochastic models. This is true not only for DSD but also Thurstonian scaling (Thurstone, 1927) and signal detection theory (Green & Swets, 1966; Wallsten & González-Vallejo, 1994). In terms of DSD and 2DSD one source of trial variability, perhaps due to lapses in attention or variability in memory processes (Pleskac, Dougherty, Rivadeneira, & Wallsten, 2009), is in the quality of the evidence accumulated or the drift rate δ. In particular, we assume that the drift rate for stimuli in the same experimental condition is normally distributed with a mean ν 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 (Hurdle 6; Ratcliff & Rouder, 1998).

A second source of trial variability is in the starting point (z). This assumption captures the idea that perhaps due to sequential effects from trial to trial judges do not start their evidence accu-

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

mulation at the same state of evidence. In DSD and 2DSD this variability is modeled with a uniform distribution mean z and range s_z. Variability in the starting point across trials leads DSD to predict faster decision times for incorrect choices than for correct choices, especially when speed is emphasized (Ratcliff & Rouder, 1998; Ratcliff et al., 1999). Although this pattern is not strongly present in our data, we have also included this assumption of trial variability in the starting point when fitting the 2DSD interrogation model.

A complete description of the 2DSD interrogation model parameters is given in Table 2. For each decision task, we fit a highly constrained model to each participant's data. Except where explicitly noted, all parameters were fixed across conditions. Furthermore, no systematic response bias was assumed so that the distribution of starting points was centered at $z = 0$ and the choice thresholds were set equal to $\theta = \theta_A = \theta_B$ (recall the lower threshold is then $-\theta$). Across the different levels of difficulty only the mean drift rate ν was allowed to vary. Thus, for each participant there were five mean drift rates ν for the line length task and six νs in the city population task. For both decision tasks only the choice thresholds θ were allowed to vary between the two time pressure conditions. As an estimate of τ we used the observed mean interjudgment time for the particular condition, corrected for any non-judgment-related processing time t_{EJ}. With six confidence ratings (.50, .60, .70, .80, .90, and 1.00) there are five confidence criteria. The confidence criteria were also fixed across experimental conditions and were symmetrical for R_A or R_B response (e.g., $c_{R_B,k} = -c_{R_A,k}$). Thus, in the line length task the 2DSD interrogation model had a total of 16 free parameters to account for 710 data points per participant. In the city population task there were a total of 17 free parameters to account for 852 data points per participant.[11] The estimated parameters for each person and task are shown in Table 6.

To evaluate the fit of the 2DSD interrogation model we adapted the latency-probability function plots first used by Audley and Pike (1965) to include confidence as well. We call the plots latency–confidence–choice probability (LCC) functions. Figure 2 shows the plots for the average participant during the line length task (top row) and the corresponding plots for the city population task (bottom row). Overall each LCC function can be understood as a plot showing how the measures of cognitive performance change with stimulus discriminability. Within each figure, the lower half of the figure plots the mean decision times against the choice probability for correct choices (gray circles right of choice probability .5) and incorrect choices (white circles left of choice probability .5) for each level of difficulty. Because the sequential sampling models were fit to five levels of difficulty in the line length task there are five correct (gray) and five incorrect (white) circles. In the city population task there are six correct (gray) and six incorrect (white) dots. The gray dot furthest to the right in each panel corresponds to the choice proportion and corresponding decision time for the easiest condition. In other words, it has the highest choice probability and typically the fastest decision time among corrects. The gray dot closest to the .5 choice probability corresponds to the proportion correct in the most difficult condition and the slowest average decision time among corrects. The white dots are configured similarly for incorrect choices with the easiest condition on the far left. The upper portion of each panel plots the mean confidence against the choice probability in the same fashion. The solid lines marked with squares correspond to the predicted functions for the 2DSD interrogation model. Recall that although the LCC functions plot the means, the models were fit using the quantiles of the decision times.

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

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

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

Trial variability in 2DSD also helps account for other phenomena. In particular, trial variability in the drift rate helps account for the observed relationship between confidence and decision times even when stimulus difficulty is held constant. Past studies (Baranski & Petrusic, 1998; Henmon, 1911) have indicated that even when the stimulus difficulty is held constant (i.e., responding to repeated trials of the same line pairs) there is a negative relationship between decision time and confidence so that the fastest

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

Table 6
Parameter Estimates for Each Participant From the 2DSD Interrogation Model With Trial Variability in the Drift Rate and Starting Point

	Participants in line length task							Participants in city population task						
Parameter	1	2	3	4	5	6	*M*	1	2	3	4	5	6	*M*
ν_1	.0521	.0375	.0261	.0322	.0372	.0389	.0373	.0118	.0239	.0205	.0181	.0170	.0158	.0179
ν_2	.1267	.1032	.0759	.0781	.1112	.0871	.0970	.0234	.0456	.0430	.0435	.0533	.0346	.0406
ν_3	.2171	.1961	.1418	.1661	.1969	.1544	.1787	.0276	.0739	.0627	.0809	.0697	.0554	.0617
ν_4	.2762	.2578	.1890	.2308	.2472	.2473	.2414	.0506	.0989	.0757	.1036	.0977	.0856	.0854
ν_5	.3527	.3719	.2655	.3446	.3189	.3762	.3383	.0565	.1497	.0968	.1477	.1259	.1348	.1186
ν_6								.0952	.2264	.1476	.2070	.2168	.1991	.1820
η	.1477	.1757	.0876	.1292	.0930	.1583	.1319	.0484	.1295	.0662	.1085	.1419	.1532	.1080
θ_{speed}	.0634	.0490	.0455	.0463	.0536	.0493	.0512	.0679	.0674	.0857	.0699	.0834	.0741	.0747
θ_{acc}	.0854	.0770	.0932	.1722	.0797	.1440	.1086	.0908	.1141	.1570	.2125	.2013	.1958	.1619
s_z	.0580	.0281	.0110	.0378	.0312	.0000	.0277	.0430	.0391	.0000	.0440	.0411	.0000	.0279
t_{ED}	.3234	.3747	.2920	.4084	.3240	.3966	.3532	.4510	.7847	.4590	.7289	.5579	.7709	.6254
t_{EJ}	.0660	.0000	.0009	.0000	.0000	.0000	.0112	.0000	.0000	.0000	.0000	.0000	.0000	.0000
c_1	.0750	−.0143	−.0713	−.0956	−.0193	.0706	−.0092	.1325	.0210	−.0199	−.1075	.1147	.0809	.0370
c_2	.1066	.0543	−.0024	−.0068	−.0193	.1501	.0471	.1714	.0967	.0518	.0179	.1570	.2226	.1196
c_3	.1157	.1251	.0491	.0509	−.0193	.2151	.0894	.1851	.1553	.1061	.1370	.1936	.3256	.1838
c_4	.1241	.1739	.0691	.1038	−.0193	.2861	.1230	.1942	.2042	.1817	.2255	.2274	.4275	.2434
c_5	.1482	.2125	.1095	.2126	−.0145	.3595	.1713	.1998	.2499	.2140	.2877	.2506	.5386	.2901

Note. 2DSD = two-stage dynamic signal detection; ν = mean value of the drift rate between trials; η = standard deviation of the drift rate between trials; θ_{speed} and θ_{acc} = choice thresholds for speed and accuracy, respectively; s_z = range of starting points; t_{ED} = mean nondecision time; t_{EJ} = mean nonjudgment time; c_1–c_5 = confidence criteria.

decision times are generally associated with the highest confidence rating. Indeed participants in the line length task tended to show this pattern of results. Table 7 lists the average Goodman and Kruskal Γ rank–order correlation between decision time and confidence for each participant in the line length task holding stimulus difficulty constant. These correlations reveal that for nearly everyone, holding stimulus difficulty constant, there was a negative correlation between decision time and confidence and the strength of this relationship was strongest for correct responses during the accuracy conditions.

The 2DSD model (without trial variability) predicts that for a given stimulus, decision times and confidence are independent of each other (separately for correct and incorrect choices). The 2DSD model with trial variability attributes this relationship between confidence and decision time to trial-by-trial variability in the drift rate. That is, even when stimulus discriminability is held constant there is a negative relationship between decision time and confidence (Hurdle 4 in Table 1) because from trial to trial there are fluctuations in the processing of the same stimuli that lead to changes in the quality of the evidence being accumulated. On some trials when the quality of the evidence extracted from a stimulus is high the drift rate will be high and lead to fast decision times and high levels of confidence. On other trials when the quality of the evidence is low the drift rate will be low and lead to slow decision times and lower levels of confidence. The model does, however, underestimate the relationship between observed decision time and confidence. For example, as listed in Table 7, the average observed Goodman and Kruskal Γs between decision time and confidence across participants and difficulty levels were –.21 and –.39 for correct choices in speed and accuracy conditions, respectively. The average predicted correlations generated from the best fitting model parameters for each participant for the speed and accuracy conditions were –.10 and –.15, respectively. Nevertheless, this result provides further support for the necessity of including trial variability in the process parameters to give a full account of the data.

Accuracy of Confidence Ratings

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

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

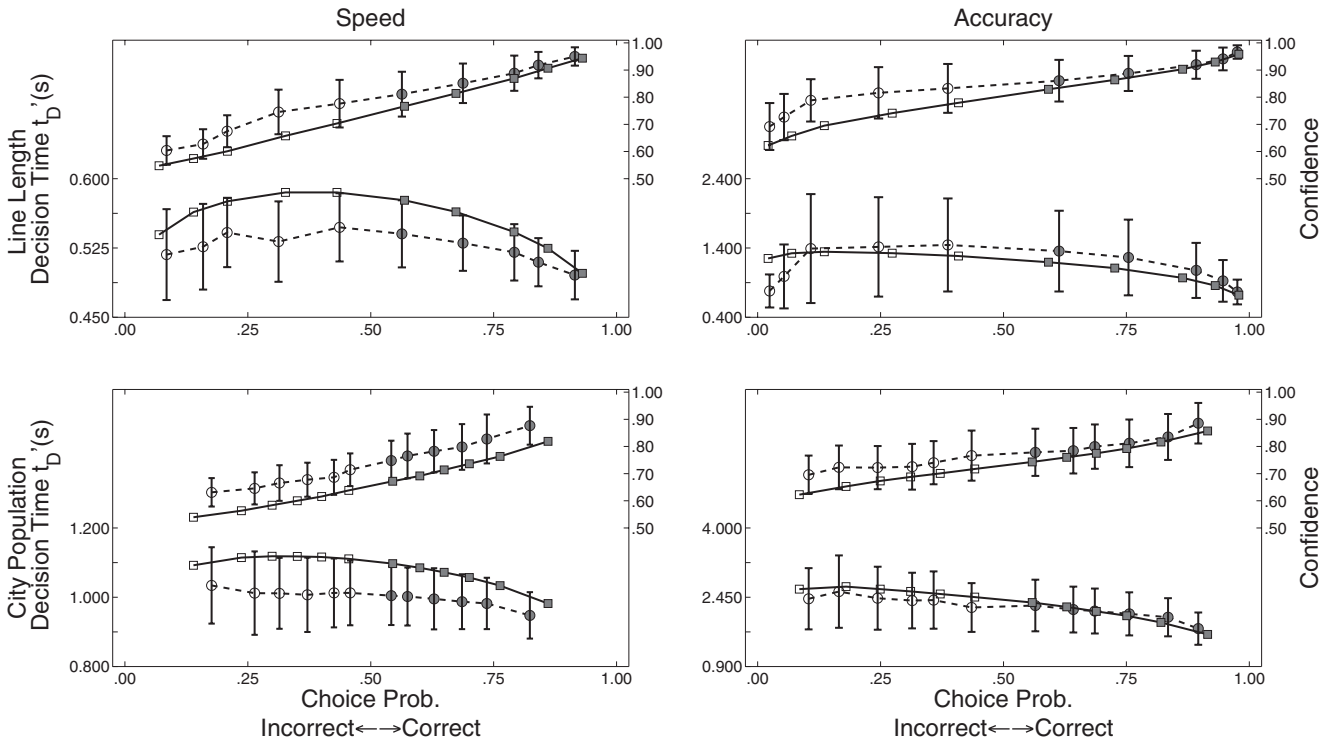

Figure 2. The latency–confidence–choice probability functions for the average participant during the line length (top row) and city population (bottom row) discrimination tasks. The best fitting functions for the two-stage dynamic signal detection (2DSD) interrogation model with trial variability in the starting point and drift rate are shown. The circles with dashed lines are the data, and the squares with solid lines are the fits of the 2DSD model. Unshaded markers are the error or incorrect responses; shaded markers are the correct responses. The error bars represent ±1.96 *SE*.

Substantive goodness. There are, in fact, two different dimensions by which the accuracy of subjective probability estimate can be evaluated: substantively and normatively (Winkler & Murphy, 1968). *Substantive goodness* captures the idea that forecasters should be able to distinguish between events that occur or not with their confidence estimates (*resolution*). In other words, do confidence ratings reflect whether the judge has made a correct choice?

Table 7
Average Goodman and Kruskal Γ *Between Confidence and Decision Time Holding Difficulty Constant for the Line Length Task for Each Participant*

	Participant					
Condition	1	2	3	4	6	*M*
Speed						
Correct	−.25*	−.34*	.01	−.21*	−.29*	−.21*
Incorrect	−.08	−.27*	.07	.11	−.15*	−.07
Accuracy						
Correct	−.35*	−.33*	−.30*	−.54*	−.50*	−.39*
Incorrect	−.12	−.25*	.02	−.21	−.30*	−.18*

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

The slope (difference between the mean confidence rating for correct choices and the mean confidence rating for incorrect choices) is one measure of substantive goodness (Yates, 1990; Yates & Curley, 1985). As we showed in the present study, the slope scores increased under time pressure in both perceptual and general knowledge tasks (see Table 5). This increase in slope is indicative of a resolution in confidence about the accuracy of their decision under time pressure at choice. 2DSD attributes this enhanced resolution to the increase in evidence collection, whether it is with an increase in interjudgment times τ in the interrogation model or a decrease in exit probabilities w in the optional stopping model (the optional stopping model is described in full later).

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

$$DI' = \frac{slope}{\sqrt{scatter}}. \quad (18)$$

Table 8 lists the means and standard deviations of the DI' scores across participants in each task (individual estimates can calcu-

Table 8
DI′, Bias, and Brier Scores Across Participants in the Speed and Accuracy Conditions of Each Task

	Line length			City population		
Variable	*M*	*SE*	$SD_{between}$	*M*	*SE*	$SD_{between}$
DI′						
Speed	1.32	0.08	0.19	0.83	0.10	0.23
Accuracy	0.99*	0.06	0.16	0.60*	0.04	0.09
Bias						
Speed	.08	.03	.07	.10	.03	.06
Accuracy	.05*	.02	.06	.07*	.03	.07
Brier						
Speed	.141	.005	.012	.209	.005	.013
Accuracy	.113*	.006	.014	.191*	.011	.026

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

lated using the values in Table 5). Using *DI′* as an index of resolution, we still see an increase in resolution during the speed conditions of both tasks.[12] Baranski and Petrusic (1994) reported a similar result. This increase in standardized resolution is best understood within the context of the 2DSD interrogation model. Recall in general that in DSD standardized accuracy grows as a linear function of the square root of time (e.g., τ; see Equation 5). This finding (increased resolution in confidence judgments when facing time pressure at choice but not during confidence rating) is added as the eighth and final empirical hurdle any model must explain (see Table 1). Race models using a one-stage balance of evidence hypothesis do not clear this empirical hurdle.

Normative goodness. The second dimension of the accuracy of subjective probabilities is *normative goodness.* Normative goodness addresses the idea that when confidence ratings come in the form of subjective probabilities, we also demand them to adhere to the properties of probabilities. This adherence is because decision makers use subjective probability judgments, like the likelihood of rain tomorrow or the probability that a sports team will win, to weigh the costs and benefits of different outcomes in making a variety of decisions. Thus, confidence ratings when given as subjective probabilities should also be evaluated in terms of their normative goodness or how well they meet the demands of probabilities (Winkler & Murphy, 1968). We can further break normative goodness into *coherence* and *correspondence.* The first factor of normative goodness is coherence or the degree to which estimates conform to the mathematical properties of probabilities specified in the Kolmogorov axioms of probability (see e.g., Rottenstreich & Tversky, 1997; Tversky & Kahneman, 1974; Tversky & Koehler, 1994). For this article 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 he or she is correct is 75% then is he or she actually correct 75% of the time? Note that correspondence in subjective probability estimates implies coherence, but coherence does not imply correspondence. However, correspondence does not necessarily imply good resolution or substantive goodness. For example, a weather forecaster who uses the long-run historical relative frequency of rain during a particular month as his or 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 tasks. One measure of correspondence is the difference between the average confidence rating and the proportion correct,

$$bias = \overline{conf} - pc, \tag{19}$$

where $bias > 0$ indicates overconfidence.[13] That is, judges tend to overestimate the likelihood they are correct. Table 8 lists the means and standard deviations of the bias scores across participants in the speed and accuracy conditions of both tasks (individual estimates can calculated using the values in Table 3). The bias scores show that most participants were on average overconfident in both the line length and city population tasks. Past results have sometimes found underconfidence in perceptual tasks like the line length task and overconfidence 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 return to this two versus one process argument in the discussion, but note that by and large 2DSD has provided an adequate account of the data in both tasks, suggesting perhaps that the distinction between perceptual and conceptual tasks is more a difference in information rather than a difference in process.

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

Reliability diagrams or calibration curves are a useful way to visualize the correspondence between stated confidence ratings and the proportion of correct inferences. Calibration curves plot the relative frequencies of an event occurring (correct choice) for given respective discrete forecasts (.50, .60, . . . , 1.00; Lichtenstein et al., 1982; Murphy & Winkler, 1977). Figure 3 shows the calibration plot for the two different tasks for Participant 2. The

[12] This increase in resolution in the speed conditions is evident even with Goodman and Kruskal's ordinal measure of association Γ (see Table 4; see Nelson, 1984, 1987, for an argument as to the use of Γ as a primary measure of accuracy).

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

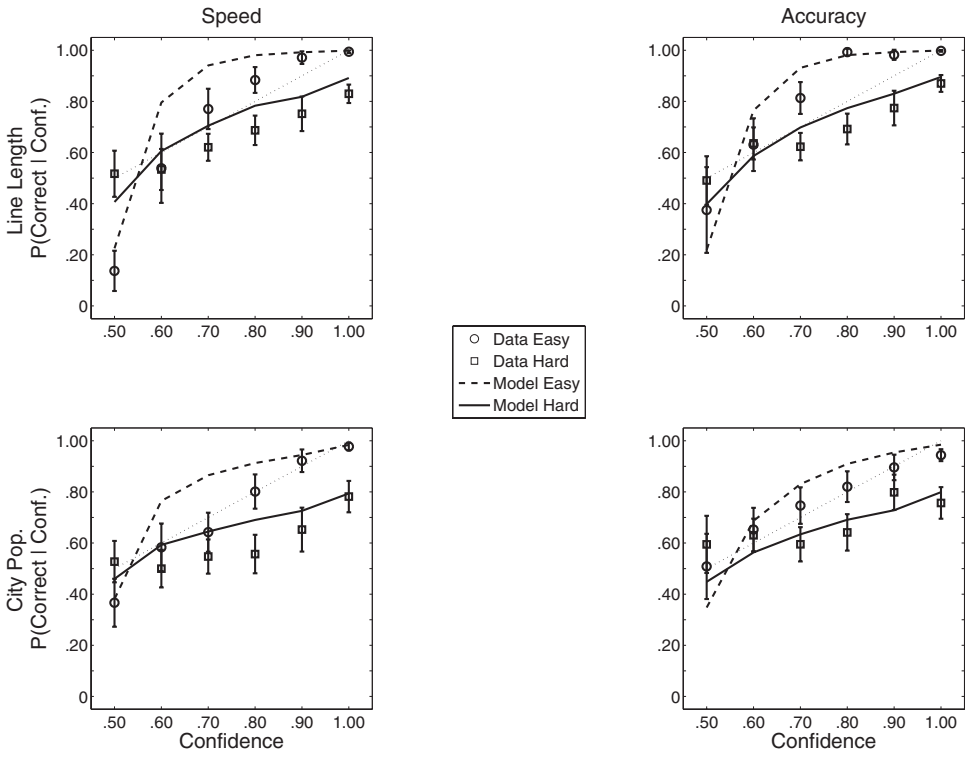

Figure 3. Empirical and best fitting (model) calibration curves for Participant 2 in the line length (top row) and city population (bottom row) discrimination tasks. The easy condition is the easiest three levels and the hard condition is the hardest three levels in the respective tasks. The error bars represent 95% confidence intervals calculated using the standard error of the proportion correct conditional on the confidence rating category.

hard condition represents the hardest three conditions of each task and the easy condition represents the easiest three. The plotted lines represent the values calculated using the 2DSD interrogation model with trial variability. The parameters that were used were the best fitting parameters from fitting the model to the distributions of the cognitive performance indices. The figure shows that the 2DSD model does a pretty good job of capturing the changes in calibration across the various conditions. The mean absolute deviation between the observed and predicted calibration curves for this participant was .08 in both the line length and city population tasks. The mean absolute deviation across all participants was .09 for both tasks.

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

$$bias = pc \cdot (slope - 1) + \overline{conf}_{incorrect}. \quad (20)$$

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

As is evident in the calibration curves in Figure 3, in both tasks there was what has been called the *hard–easy effect* where judges tend to grow less overconfident as choices get easier and even exhibit underconfidence for the easiest questions (Ferrell & McGoey, 1980; Gigerenzer et al., 1991; Griffin & Tversky, 1992; Juslin, Winman, & Olsson, 2000; Lichtenstein et al., 1982). In the line length task, collapsing across the easiest three conditions, across participants the bias score went from an average of .13 ($SE = .004$; $SD_{between} = .09$) in the hardest three conditions to .00 ($SE = .002$; $SD_{between} = .05$) in the easiest conditions. In the city population task, these numbers were .15 ($SE = .005$; $SD_{between} = .07$) and .02 ($SE = .004$; $SD_{between} = .06$), 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 (ν 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 et al., 1996).

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

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

$$brier = (conf_i - correct_i)^2. \quad (21)$$

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 (.50, .60, . . . , 1.00). In this case the goal of the judge is to produce judgments that minimize the Brier score. The Brier score is a useful overall measure of accuracy for two reasons. One reason is that the Brier score is a strictly proper scoring rule (Aczel & Pfanzagal, 1966; de Finetti, 1962; Murphy, 1973; von Winterfeldt & Edwards, 1986; Yates, 1990). This means that if participants have an explicit goal of minimizing their Brier score they will achieve this minimum if they (a) give an accurate choice and (b) truthfully map their internal subjective belief that they are correct, p^*, to the closest external probability value available (*conf*). Recall that participants were rewarded in both tasks according to a linear transformation of the Brier score (see Equation 15). Thus, it was in participants' best interests to have a goal of minimizing their Brier score.

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

$$\overline{brier} = VI + bias^2 + (VI)(slope)(slope - 2) + scatter, \quad (22)$$

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

Table 8 lists the average Brier score under speed and accuracy instructions for both the line length and city population discrimination tasks. The standard error for the average Brier score was estimated using bootstrap methods. By and large, participants tended to have worse (higher) 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 .211 (SE = .065; $SD_{between}$ = .021) in the hardest three conditions, which was significantly larger than the average Brier score in the easiest three conditions, .045 (SE = .019; $SD_{between}$ = .006). In the city population task, the average Brier score was .260 (SE = .029; $SD_{between}$ = .019) in the hardest three conditions, which was significantly larger than the average Brier score in the easiest three conditions, .142 (SE = .023; $SD_{between}$ = .033).

Using the Brier score decomposition in Equation 22 we can see how changes in different types of accuracy lead to changes in the Brier score. For example, a decrease in choice accuracy will by and large increase the mean Brier score via the variability index. But, an increase in slope will decrease the Brier score (the derivative of Equation 22 with respect to slope is negative). According to 2DSD when judges increased the amount of evidence collected during the second stage of processing under time pressure at choice they were at least in principle counteracting the negative impact of lower choice accuracy on the Brier score and their final earnings. To investigate this claim we used the 2DSD interrogation model with trial variability to evaluate how the Brier score changes as a function of (a) choice thresholds θ and (b) interjudgment times τ. Figure 4 plots this relationship for Participant 3 in the city population task. It shows that according to 2DSD the Brier score will decrease when choice thresholds are increased and/or when interjudgment times τ are increased.[14]

But of course this change in choice thresholds and interjudgment times exacts some costs on the judge. In particular, the improvement in overall accuracy comes at the cost of greater processing time. A way to conceptualize these different demands of accuracy and time is with the cost function

$$cost = c_1 \cdot t'_D + c_2 \cdot \tau' + c_3 \cdot brier, \quad (23)$$

where c_1, c_2, and c_3 specify the cost of the observed decision time t'_D, the observed interjudgment time τ', and the level of a Brier score, respectively.

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

[14] Note this relationship with the Brier score is dependent to some degree on the location of the confidence criteria/markers and drift rate. For example, at extreme levels of drift rate/mean drift rate, changes in interjudgment time and choice thresholds have little effect on the Brier score. Furthermore, the relationship also breaks down when judges use what appear to be atypical locations of confidence criteria/markers (e.g., only using the .50, .90, and 1.00 confidence ratings).

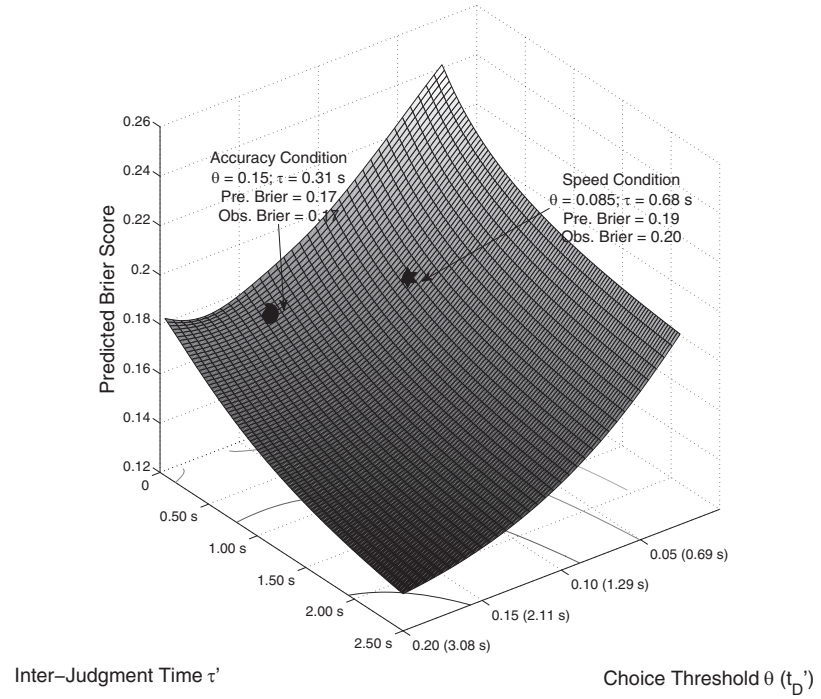

Figure 4. Predicted Brier scores for Participant 3 in the city population task in Difficulty Level 3. The plot was calculated using the two-stage dynamic signal detection (2DSD) interrogation model with trial variability in the parameters. The plot illustrates that according to the 2DSD model, increases in the choice threshold θ and interjudgment time τ both minimize a person's Brier score. This implies that the model can be used to find appropriate choice threshold and interjudgment time settings that produce the fastest total judgment time (choice + confidence) for a given Brier score.

2DSD Optional Stopping Model Account of Interjudgment Times

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

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

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

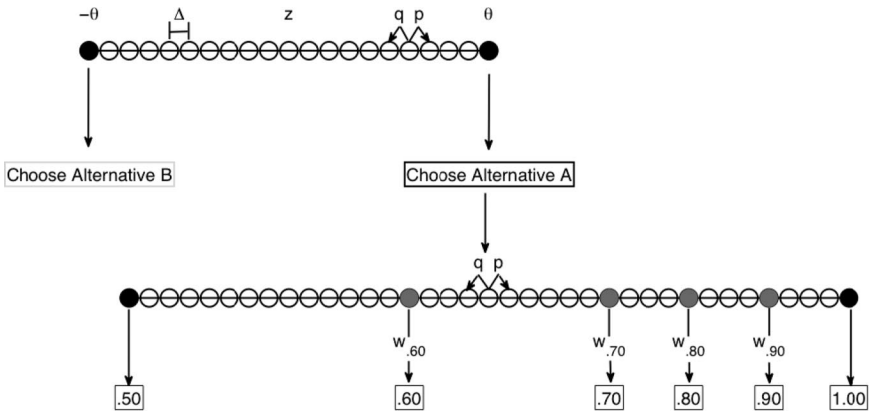

Figure 5. A Markov chain approximation of the two-stage dynamic signal detection (2DSD) optional stopping model. In the model, evidence accumulates over time toward an upper (θ) and lower (−θ) threshold. This accumulation is approximated with discrete states in the model using probabilities p and q of moving a step size Δ to each adjacent state. This process can produce a trajectory such as the jagged line in Figure 1. After making a choice, judges continue accumulating evidence but are assumed to lay out markers κ_j across the state space so that if the process crosses through that particular state, judges exit with probability w_j and give the corresponding confidence rating. These markers are identified with gray and black dots in the bottom chain. To adequately fit the data, the model assumed that the two extreme confidence ratings (e.g., .50 and 1.00) were associated with an absorbing boundary so that if the process entered its associated state the judge would stop accumulating evidence and state the respective level of confidence. If Alternative B was chosen a similar chain is used (not shown), which is a reflection of the chain used if Alternative A was chosen. The model predicts at the distribution level choice, decision time, confidence, interjudgment times, and their interrelationships (see Appendix C). z = starting point.

ities and decision times that closely approximate a continuous time diffusion process (see Appendix C; Diederich & Busemeyer, 2003).

More importantly for our interests, the discrete state space gives another means to conceptualize our two-stage hypothesis. Under this formulation, the confidence stage is modeled as a second Markov chain (see the bottom chain in Figure 4 when Alternative A is chosen). Now during the second stage of processing, we assume markers κ_j are placed along the evidence state space representing the different confidence ratings (j = .50, .60, . . . , 1.00), one for each rating. For the intermediary confidence ratings (.60, .70, .80, and .90), each time evidence passes one of these markers there is a probability w_{conf} that the judge exits and gives the corresponding confidence rating.[15] The evidence states representing the confidence ratings of .50 and 1.00 were set equal to an absorbing boundary ($w_{.50} = w_{1.00} = 1.0$, thus the black circles shown in the lower chain in Figure 5). This means that once the process enters one of these extreme states then with probability 1 the evidence accumulation process ends and the corresponding confidence rating is given. Using the same Markov chain methods that determine the choice and decision times, the distribution of confidence ratings and distribution of interjudgment times can be computed (see Appendix C).

Fitting the 2DSD optional stopping model. Evaluating the 2DSD optional stopping model is very challenging because it requires simultaneously fitting the entire distribution of responses across choices, decision times, confidence ratings, and interjudgment times for the different conditions of each decision task. Appendix C describes the detailed procedures used to accomplish the fits. In this section we summarize the most important points. To summarize the procedure, we fit quantiles averaged across individuals, and this fit was done separately for each of the two conditions of time pressure and decision task conditions.[16,17] Each time pressure and decision task condition had approximately 81 free data points, and the 2DSD optional stopping model had 14 free parameters. The parameter estimates are shown in Table 8.[18]

The most important finding that needs to be explained with this 2DSD optional stopping model is the relation between confidence ratings and interjudgment times. Figure 6 displays the quantiles (.1, .3, .5, .7, and .9) of the distribution of interjudgment times as a function of confidence and accuracy, averaged across all six

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

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

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

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

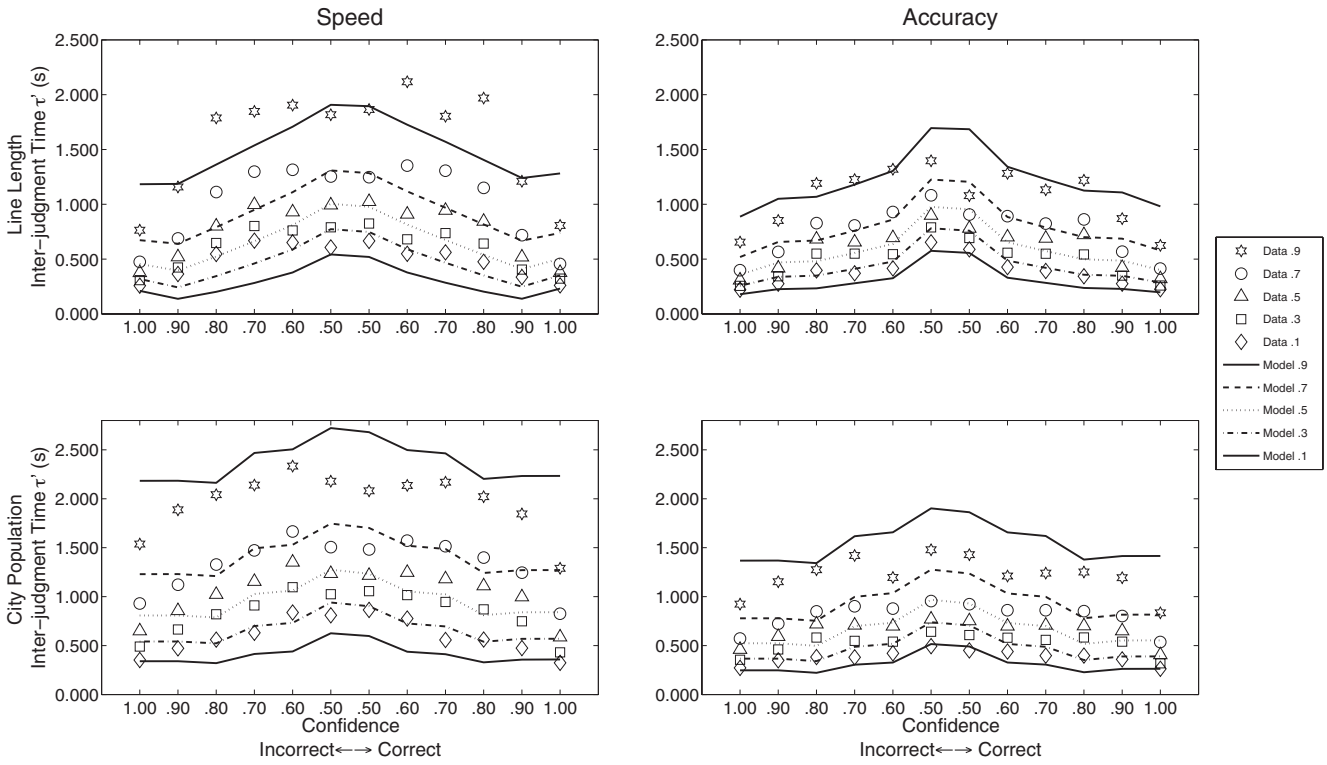

Figure 6. Observed and best fitting (model) distribution of interjudgment times (τ) as a function of confidence and accuracy in the line length and city population tasks for the average participant. The observed and best fitting distributions are plotted using the .1, .3, .5, .7, and .9 quantiles of the respective distributions. The best fitting distribution of interjudgment times was generated from the 2DSD optional stopping model formulated as a Markov chain. As the figures show, the interjudgment times for both corrects and incorrects on average grew faster with increasing levels of confidence, and the 2DSD model by and large gives a good account of the data.

participants, separately for each decision task and speed instruction condition. In particular, the triangles display the median interjudgment time, which is largest when confidence is lowest and decreases as confidence becomes more extreme.

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

Table 9
Parameter Estimates for the 2DSD Optional Stopping Model Fit to the Average Choice, Decision Time, Confidence, and Interjudgment Time Distributions

Task	δ	θ	t_{ED}	$w_{.60}$	$w_{.70}$	$w_{.80}$	$w_{.90}$	$\kappa_{.50}$	$\kappa_{.60}$	$\kappa_{.70}$	$\kappa_{.80}$	$\kappa_{.90}$	$\kappa_{1.00}$	t_{EJ}
Line length														
Speed	.0724	.0709	.1750	.0154	.0107	.0075	.0092	−.1000	−.0273	.0014	.0264	.0480	.1457	.0785
Accuracy	.0776	.1088	.1466	.0319	.0291	.0178	.0280	−.0870	.0116	.0283	.0463	.0495	.1673	.0936
City population														
Speed	.0329	.0958	.2821	.0137	.0138	.0251	.1075	−.0686	.0076	.0119	.1671	.1922	.1968	.1053
Accuracy	.0335	.1409	.2894	.0214	.0204	.0250	.1185	−.0111	.0704	.0745	.1883	.2168	.2206	.1076

Note. 2DSD = two-stage dynamic signal detection; δ = drift rate; θ = choice threshold; t_{ED} = mean nondecision time; w = exit probability; κ = confidence rating marker; t_{EJ} = mean nonjudgment time.

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

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

Discussion

We have recast the standard diffusion model to give a single process account of choice, decision time, and confidence. Perhaps the best summary of 2DSD is provided by the answer to a recent question from a colleague curious whether "choice, response [decision] time, or confidence was a purer measure of the judgment process during a two-alternative forced choice task?" 2DSD implies that no one dependent variable is a pure measure of the judgment process. Instead these three measures simply give different views of the same underlying evidence accumulation process.

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

Cognitive Underpinnings of Confidence

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

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

An Interaction Between Choice and Confidence Stages

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

This interaction between choice and confidence stages is difficult for most other models of confidence to explain. Of course the results automatically rule out signal detection models (Green & Swets, 1966), which are silent on the time course of choice and confidence judgments. Other sequential sampling models like the Poisson race model (Merkle & Van Zandt, 2006; Van Zandt, 2000b), the linear ballistic accumulator (Brown & Heathcote, 2008), or the leaky accumulator (Usher & McClelland, 2001,

2004), which use the balance of evidence hypothesis (based solely on evidence at the time of choice), also cannot easily handle this pattern. As we indicated earlier, race models, however, may be able to account for the interaction between choice and confidence stage if they are given a second stage of evidence accumulation. That is, after making a choice the counters continue racing to a second confidence threshold and confidence is calculated according to the balance of evidence at this second threshold (see e.g., Van Zandt & Maldonado-Molina, 2004). Future comparisons between two-stage race models and 2DSD will certainly be revealing and better characterize the interaction between the two stages.

This interaction between choice and confidence stages also appears to be difficult for other confidence models to qualitatively handle. Consider for example Ratcliff and Starns's (2009) recent model of response time and confidence, RTCON. RTCON is restricted to no-choice tasks where judges rate only their confidence, say, from 0% (certain a prespecified item is incorrect) to 100% (certain a prespecified 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 assumed that a stimulus item at test produces some degree of activation that 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 showed that the RTCON model can explain several different phenomena related to estimated ROC functions from confidence ratings in recognition memory tasks.

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

Possible Models of the No-Choice Task

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

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

Stopping Rules of the Second Stage

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

The data, however, also suggest that the interjudgment times may be determined by the judgment system itself. In other words, interjudgment times may be an endogenous variable. The 2DSD optional stopping model captures this idea where during the second stage of evidence accumulation markers are spread throughout the space. These markers are in turn used to select a confidence rating based on the location of the accumulated evidence. This model not only accounts for the three primary measures of cognitive performance but also gives a good account of the distribution of interjudgment times. By and large, these two models mimic each other in terms of the predictions of choice, decision time, and confidence and the effects of the different conditions on these variables. Moreover, although both models are mathematically feasible, the 2DSD interrogation model due to its continuous

nature is computationally easier to fit. These properties make the 2DSD interrogation model advantageous for a number of applications seeking to understand cognitive performance via measures of choice, decision time, and confidence. Examples of applications include the instances from the decision sciences (e.g., Arkes, 2001), human factors (e.g., Sanders & McCormick, 2002), and any other applied setting where confidence/subjective probabilities are a crucial measure of one's belief that an event has occurred or will occur (de Finetti, 1962).

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

A Common Choice and Judgment Process

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

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

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

Accuracy of Confidence

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

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

More generally, this pattern of findings demonstrates that without understanding the goals of judges in our study—and the role of accuracy within these goals—we would not understand the observed behavior of judges. At the same time, though, the increase in interjudgment time does not make sense unless we understand the process underlying choice and confidence in terms of 2DSD. This addresses a larger issue in judgment and decision making where there has been a call for basic judgment research to orient away from questions of response accuracy and instead focus more on the process (Erev et al., 1994; Wallsten, 1996). Our results actually speak to a broader call for theories of judgment and decision making to not focus solely on process *or* accuracy. Rather we must use and understand both process *and* accuracy (and more generally judges' goals) in tandem to explain choice *and* judgment (cf. Anderson, 1990).

Conclusion

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

References

Aczel, J., & 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,* 432–436.

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

Anderson, J. R. (1990). *The adaptive character of thought.* Hillsdale, NJ: Erlbaum.

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

Arkes, H. R. (2001). Overconfidence in judgmental forecasting. In J. S. Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners* (pp. 157–176). Boston, MA: Kluwer Academic.

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

Ashby, F. G. (2000). A stochastic version of general recognition theory. *Journal of Mathematical Psychology, 44,* 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,* 207–225.

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

Baranski, J. V., & Petrusic, W. M. (1995). On the calibration of knowledge and perception. *Canadian Journal of Experimental Psychology, 49,* 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,* 929–945.

Barnard, G. A. (1946). Sequential tests in industrial statistics. *Journal of the Royal Statistical Society Supplement, 8,* 1–26.

Bhat, U. N. (1984). *Elements of applied stochastic processes.* New York, NY: Wiley.

Björkman, M., Juslin, P., & Winman, A. (1993). Realism of confidence in sensory discrimination—The underconfidence phenomenon. *Perception & Psychophysics, 54,* 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,* 700–765.

Bogacz, R., Wagenmakers, E. J., Forstmann, B. U., & Nieuwenhuis, S. (2010). The neural basis of the speed–accuracy tradeoff. *Trends in Neurosciences, 33,* 10–16.

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

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology, 57,* 153–178.

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,* 157–171.

Budescu, D. V., & Wallsten, T. S. (1995). Processing linguistic probabilities: General principles and empirical evidence. In J. R. Busemeyer, R. Hestie, & D. Medin (Eds.), *Decision making from the perspective of cognitive psychology* (pp. 275–318). New York, NY: Academic Press.

Budescu, D. V., Wallsten, T. S., & Au, W. T. (1997). On the importance of random error in the study of probability judgment: Part II. Applying the stochastic judgment model to detect systematic trends. *Journal of Behavioral Decision Making, 10,* 173–188.

Busemeyer, J. R., & Diederich, A. (2010). *Cognitive modeling.* New York, NY: Sage.

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

Busemeyer, J. R., Jessup, R. K., Johnson, J. G., & Townsend, J. T. (2006). Building bridges between neural models and complex decision making behaviour. *Neural Networks, 19,* 1047–1058.

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,* 432–459.

Busey, T. A., Tunnicliff, J., Loftus, G. R., & Loftus, E. F. (2000). Accounts of the confidence–accuracy relation in recognition memory. *Psychonomic Bulletin & Review, 7,* 26–48.

Cox, D. R., & Miller, H. D. (1965). *The theory of stochastic processes.* New York, NY: Chapman and Hall.

Dawes, R. M. (1980). Confidence in intellectual judgments vs. confidence in perceptual judgments. In E. D. Lantermann & H. Feger (Eds.),

Similarity and choice: Papers in honour of Clyde Coombs (pp. 327–345). Bern, Switzerland: Humber.

de Finetti, B. (1962). Does it make sense to speak of 'good probability appraisers'? In I. J. Good (Ed.), *The scientist speculates: An anthology of partly-baked ideas* (pp. 357–364). New York, NY: Basic Books.

Diederich, A. (1997). Dynamic stochastic models for decision making under time constraints. *Journal of Mathematical Psychology, 41,* 260–274.

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,* 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,* 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,* 312–329.

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

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review, 101,* 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 & Psychophysics, 1,* 369–373.

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

Ferrell, W. R., & McGoey, P. J. (1980). A model of calibration for subjective probabilities. *Organizational Behavior and Human Decision Processes, 26,* 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—Personal observations on the reliability of the shuttle* [Appendix to the Rogers Commission Report]. Retrieved from http://science.ksc.nasa.gov/shuttle/missions/51-l/docs/rogers-commission/Appendix-F.txt

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

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

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

Gold, J. I., & Shadlen, M. N. (2002). Banburismus and the brain: Decoding the relationship between sensory stimuli, decisions, and reward. *Neuron, 36,* 299–308.

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,* 75–90.

Gomez, P., Perea, M., & Ratcliff, R. (2007). A model of the go/no-go task. *Journal of Experimental Psychology: General, 136,* 389–413.

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

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association, 49,* 732–769.

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

Griffin, D., & Brenner, L. (2004). Perspectives on probability judgment calibration. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment & decision making* (pp. 177–199). Oxford, England: Blackwell.

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

Hanes, D. P., & Schall, J. D. (1996, October 18). Neural control of voluntary movement initiation. *Science, 274,* 427–430.

Heath, R. A. (1984). Random-walk and accumulator models of psychophysical discrimination—A critical evaluation. *Perception, 13,* 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.

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,* 841–861.

Juslin, P. (1994). The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items. *Organizational Behavior and Human Decision Processes, 57,* 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,* 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,* 384–396.

Juslin, P., Winman, A., & Persson, T. (1995). Can overconfidence be used as an indicator of reconstructive rather than retrieval processes? *Cognition, 54,* 99–130.

Keren, G. (1988). On the ability of monitoring non-veridical perceptions and uncertain knowledge—Some calibration studies. *Acta Psychologica, 67,* 95–119.

Kiani, R., & Shadlen, M. N. (2009, May 8). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science, 324,* 759–764.

Kim, J. N., & Shadlen, M. N. (1999). Neural correlates of a decision in the dorsolateral prefrontal cortex of the macaque. *Nature Neuroscience, 2,* 176–185.

Koehler, D. J., Brenner, L., & Griffin, D. (2002). The calibration of expert judgment: Heuristics and biases beyond the laboratory. In T. G. Gilovich & D. Griffin (Ed.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 686–715). New York, NY: Cambridge University Press.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126,* 349–370.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 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,* 343–352.

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.* Hillsdale, NJ: Erlbaum.

Link, S. W. (2003). Confidence and random walk theory. In B. B. E. Borg

(Ed.), *Proceeding of the nineteenth annual meeting of the International Society for Psychophysics.* Stockholm, Sweden: International Society for Psychophysics.

Link, S. W., & Heath, R. A. (1975). Sequential theory of psychological discrimination. *Psychometrika, 40,* 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 guide.* New York, NY: Erlbaum.

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, England: Wiley.

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

Moreno-Bote, R. (in press). Decision confidence and uncertainty in diffusion models with partially correlated neural integrators. *Neural Computation.*

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology, 12,* 595–600.

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

Nelder, J. A., & Mead, R. (1965). A simplex-method for function minimization. *Computer Journal, 7,* 308–313.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95,* 109–133.

Nelson, T. O. (1987). The Goodman-Kruskal gamma coefficient as an alternative to signal-detection theory's measures of absolute-judgment accuracy. In E. Roskam & R. Suck (Eds.), *Progress in mathematical psychology* (pp. 299–306). New York, NY: Elsevier Science.

Nelson, T. O. (1996). Consciousness and metacognition. *American Psychologist, 51,* 102–116.

Nelson, T. O. (1997). The meta-level versus object-level distinction (and other issues) in formulations of metacognition. *American Psychologist, 52,* 179–180.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *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, 104,* 266–300.

Pachella, R. G. (1974). The interpretation of reaction time in information-processing research. In B. H. Kantowitz (Ed.), *Human information processing: Tutorials in performance and cognition* (pp. 41–82). Hillsdale, NJ: Erlbaum.

Petrusic, W. M., & Baranski, J. V. (2003). Judging confidence influences decision processing in comparative judgments. *Psychonomic Bulletin & Review, 10,* 177–183.

Pierce, C. S. (1877). Illustrations of the logic of science: The probability of induction. *The Popular Science Monthly, 12,* 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,* 28–38.

Pike, R. (1973). Response latency models for signal detection. *Psychological Review, 80,* 53–68.

Pleskac, T. J., Dougherty, M. R., Rivadeneira, A. W., & Wallsten, T. S. (2009). Random error in judgment: The contribution of encoding and retrieval processes. *Journal of Memory and Language, 60,* 165–179.

Ratcliff, R. (1978). Theory of memory retrieval. *Psychological Review, 85,* 59–108.

Ratcliff, R. (2002). A diffusion model account of reaction time and accuracy in a brightness discrimination task: Fitting real data and failing to fit fake but plausible data. *Psychonomic Bulletin & Review, 9,* 278–291.

Ratcliff, R., Gronlund, S. D., & Sheu, C. F. (1992). Testing global memory models using ROC curves. *Psychological Review, 99,* 518–535.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science, 9,* 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,* 127–140.

Ratcliff, R., & Smith, P. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review, 111,* 333–367.

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,* 438–481.

Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review, 106,* 261–300.

Resulaj, A., Kiani, R., Wolpert, D. M., & Shadlen, M. N. (2009, September 10). Changes of mind in decision-making. *Nature, 461,* 263–268.

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

Romo, R., Hernandez, A., Zainos, A., Lemus, L., & Brody, C. D. (2002). Neuronal correlates of decision-making in secondary somatosensory cortex. *Nature Neuroscience, 5,* 1217–1225.

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

Sanders, M., & McCormick, E. (1993). *Human factors in engineering and design.* New York, NY: McGraw-Hill.

Savage, L. J. (1954). *The foundations of statistics.* New York, NY: Wiley.

Schouten, J. F., & Bekker, J. A. M. (1967). Reaction time and accuracy. *Acta Psychologica, 27,* 143–153.

Shadish, W. R., & Haddock, K. C. (1994). Combining estimates of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 261–281). New York, NY: Russell 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* (pp. 103–122). 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 and Human Decision Processes, 103,* 68–83.

Smith, P. L. (1990). A note on the distribution of response times 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,* 567–593.

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

Speckman, P. L., & Rouder, J. N. (2004). A comment on Heathcote, Brown, and Mewhort's QMLE method for response time distributions. *Psychonomic Bulletin & Review, 11,* 574–576.

Squire, L. R., Wixted, J. T., & Clark, R. E. (2007). Recognition memory and the medial temporal lobe: A new perspective. *Nature Reviews Neuroscience, 8,* 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,* 251–260.

Suantak, L., Bolger, F., & Ferrell, W. R. (1996). The hard–easy effect in subjective probability calibration. *Organizational Behavior and Human Decision Processes, 67,* 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,* 16–32.

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.

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

Tversky, A., & Kahneman, D. (1974, September 27). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124–1131.

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

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review, 108,* 550–592.

Usher, M. M., & McClelland, J. L. (2004). Loss aversion and inhibition in dynamical models of multialternative choice. *Psychological Review, 111,* 757–769.

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

Van Zandt, T. (2000b). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 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,* 208–256.

Van Zandt, T., & Maldonado-Molina, M. M. (2004). Response reversals in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and 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, Germany: 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,* 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,* 179–197.

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

Voss, A., & Voss, J. (2008). A fast numerical algorithm for the estimation of diffusion model parameters. *Journal of Mathematical Psychology, 52,* 1–9.

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,* 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,* 326–339.

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

Wallsten, T. S., Bender, R. H., & Li, Y. (1999). Dissociating judgment from response processes in statement verification: The effects of experience on each component. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 96–115.

Wallsten, T. S., Budescu, D. V., Erev, I., & Diederich, A. (1997). Evaluating and combining subjective probability estimates. *Journal of Behavioral Decision Making, 10,* 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,* 176–190.

Wallsten, T. S., & González-Vallejo, C. (1994). Statement verification: A stochastic model of judgment and response. *Psychological Review, 101,* 490–504.

Wang, X. J. (2002). Probabilistic decision making by slow reverberation in cortical circuits. *Neuron, 36,* 955–968.

Wickelgren, W. A. (1977). Speed–accuracy tradeoff and information-processing dynamics. *Acta Psychologica, 41,* 67–85.

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

Winman, A., & Juslin, P. (1993). Calibration of sensory and cognitive judgments—Two different accounts. *Scandinavian Journal of Psychology, 34,* 135–148.

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

Yates, J. F., & Curley, S. P. (1985). Conditional distribution analyses of probabilistic 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,* 1341–1354.

(*Appendices follow*)

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 e.g., Busemeyer & Diederich, 2010; 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, choice threshold θ, and drift coefficient σ, the probability of choosing Alternative A, R_A, for a Wiener process is

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

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

The finishing time probability density function (PDF) for the time that the activation reaches θ and the judge responds to the given stimulus is

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

The cumulative distribution function (CDF) is

$$G(t_D|R_A,S_A) = 1 - \frac{1}{P(R_A|S_A)}\pi\left(\frac{2\theta}{\sigma}\right)^{-2}\exp\left[\frac{\delta(\theta - z)}{\sigma^2}\right]\sum_{k=1}^{\infty}\frac{2k\sin\left[\frac{k\pi(\theta - z)}{2\theta}\right]\exp\left\{\frac{-t_D}{2}\left[\frac{\delta^2}{\sigma^2} + \left(\frac{\pi k\sigma}{2\theta}\right)^2\right]\right\}}{\frac{\delta^2}{\sigma^2} + \left(\frac{\pi k\sigma}{2\theta}\right)^2}. \tag{A3}$$

The expressions for the PDF and CDF of the finishing times when stimulus S_B is present can be found by replacing δ with $-\delta$ and exchanging the choice probability. The expressions when R_B is given can be found by replacing $(\theta - z)$ with $(\theta + z)$ and again changing the choice probability.

Several items should be noted here on using Equations A1 to A3. In some models using a random walk/diffusion process it proved necessary to not assume the drift rate (or the mean drift rate) for S_B to be the negative of the drift rate of S_A, $-\delta$ (see Ratcliff, 1978; Ratcliff & Smith, 2004). A second point is that in using Equations A2 and A3 to calculate the PDF and CDF the so-called fly in the ointment is the summation to infinity. To work around this Ratcliff and Tuerlinckx (2002) recommend iteratively summing the expression within the sum until the current term and previous term are both less than 10^{-29} times the current sum (p. 478). Programming these in MATLAB, we used a fixed value of k that always met this requirement ($\sim$200). An alternative method is to use a numerical routine like Voss and Voss's (2008) fast-dm routine. Finally, the PDF and CDF in Equations A2 and A3 grow unstable as t_D gets very small. Van Zandt, Colonius, and Proctor (2000) suggested using alternative forms of the PDF and CDF for very small values.

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

$$E(t_D|R_A,S_A) = \frac{1}{\delta}\left(\frac{2\theta\left\{\exp\left[\frac{2(\theta + z)\delta}{\sigma^2}\right] + \exp\left[-\frac{2(\theta - z)\delta}{\sigma^2}\right]\right\}}{\exp\left[\frac{2(\theta + z)\delta}{\sigma^2}\right] - \exp\left[-\frac{2(\theta - z)\delta}{\sigma^2}\right]} - \frac{(\theta + z)\left\{\exp\left[\frac{2(\theta + z)\delta}{\sigma^2}\right] + 1\right\}}{\exp\left[\frac{2(\theta + z)\delta}{\sigma^2}\right] - 1}\right). \tag{A4}$$

The mean time to choose R_B when presented with S_A is

$$E(t_D|R_B,S_A) = \frac{1}{\delta}\left(\frac{2\theta\left\{\exp\left[\frac{2(\theta+z)\delta}{\sigma^2}\right]+\exp\left[-\frac{2(\theta-z)\delta}{\sigma^2}\right]\right\}}{\exp\left[\frac{2(\theta+z)\delta}{\sigma^2}\right]-\exp\left[-\frac{2(\theta-z)\delta}{\sigma^2}\right]} - \frac{(\theta-z)\left\{\exp\left[-\frac{2(\theta-z)\delta}{\sigma^2}\right]+1\right\}}{\exp\left[-\frac{2(\theta-z)\delta}{\sigma^2}\right]-1}\right). \quad \text{(A5)}$$

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

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

Appendix B

Quantile Maximum Probability Method for Fitting Decision Times and Confidence

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

The basic idea of the QMP method is that the quantile estimates form six categories of decision times. Within each category we can determine the frequency of decision times falling between the two boundaries (e.g., 20% of the responses fall within the quantiles for .1 and .3). Then using the multinomial distribution function we can calculate the likelihood of the data L (e.g., the likelihood of incorrect decision times during the speeded line discrimination task) for each model using the CDF 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 a 6 (decision time categories) $\times$ 6 (confidence ratings) data matrix for corrects and incorrects for both tasks. We fit the 2DSD interrogation model to this 6 $\times$ 6 data matrix for corrects and incorrects for both the line length and city population tasks.

Notice that the 2DSD model without trial variability predicts that for a given stimulus with a particular drift rate the distributions of decision times are independent conditional on whether the choice was correct. That is, for correct and incorrect choices for a given stimulus the distributions of decision times and confidence ratings are independent. Thus, fitting the marginal distributions will produce the same result as fitting the joint distribution of decision times and confidence ratings. In both the line length and city population tasks we broke the data down into two levels of time pressure crossed with different levels of difficulty. Recall that during the line length task, we collapsed the sixth and fifth levels of difficulty for both speed and accuracy, forming five levels of difficulty to model. In the city population task, we formed six different levels of difficulty with

(*Appendices continue*)

approximately 300 pairs in each condition. The 2DSD models were fit at the individual level to a line length task where there were 10 conditions (speed vs. accuracy $\times$ 5 levels of difficulty). Representing the data in terms of our adapted QMP method, each condition had 71 free data points, producing a total of $71 \times 10 = 710$ data points per participant. The city population task with 12 conditions (speed vs. accuracy $\times$ 6 levels of difficulty) had $71 \times 12 = 852$ free data points per participant. We fit highly constrained models to each participant's data (for a description of the constraints see the model estimation section). In total there were 16 free parameters (for 710 free data points) in the line length task and 17 free parameters (for 852 free data points) in the city population task.

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

Appendix C

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

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

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

$$L = \begin{Bmatrix} -k\Delta, & -(k-1)\Delta, & \cdots & -\Delta, & 0, & \Delta, & \cdots & (k-1)\Delta, & k\Delta \\ 1 & 2 & \cdots & & (m-1)/2 & & \cdots & m-1 & m \end{Bmatrix} \tag{C1}$$

where $\theta = k\Delta$ and $-\theta = -k\Delta$. The transition probabilities for the m states of the Markov chain are arranged in an $m \times m$ transition probability matrix $\mathbf{P}$ with the elements $p_{1,1} = 1$ and $p_{m,m} = 1$, and for $1 < i < n$,

$$P_{i,j} = \begin{cases} \dfrac{1}{2\alpha}\left\{1 - \dfrac{\delta[-k\Delta + (i-1)\Delta]}{\sigma^2}\sqrt{\rho}\right\} & \text{if } j - i = -1 \\ \dfrac{1}{2\alpha}\left\{1 + \dfrac{\delta[-k\Delta + (i-1)\Delta]}{\sigma^2}\sqrt{\rho}\right\} & \text{if } j - i = +1 \\ 0 & \text{otherwise} \end{cases} \tag{C2}$$

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 approaches zero and setting $\Delta = \alpha\sigma\sqrt{\rho}$, the random walk will converge to a Wiener diffusion process which has a continuous time set and continuous state space. The parameter $\alpha > 1$ improves the approximation of the 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 $\mathbf{P} = \| p_{i,j} \|$ is presented in its canonical form:

$$\mathbf{P} = \left| \begin{array}{c|c} \mathbf{P_1} & \mathbf{0} \\ \hline \mathbf{R} & \mathbf{Q} \end{array} \right| = \left| \begin{array}{c|cc|cccc|cc} 1 & 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ m & 0 & 1 & 0 & 0 & \cdots & 0 & 0 \\ \hline 2 & p_{2,1} & 0 & p_{2,2} & p_{2,3} & \cdots & 0 & 0 \\ 3 & 0 & 0 & p_{3,2} & p_{3,3} & \cdots & 0 & 0 \\ 4 & 0 & 0 & 0 & p_{4,3} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ m-3 & 0 & 0 & 0 & 0 & \cdots & p_{m-3,m-2} & 0 \\ m-2 & 0 & 0 & 0 & 0 & \cdots & p_{m-2,m-2} & p_{m-2,m-1} \\ m-1 & 0 & p_{m-1,m} & 0 & 0 & \cdots & p_{m-1,m-2} & p_{m-1,m-1} \end{array} \right|. \quad \text{(C3)}$$

with $\mathbf{P}_1$ being a 2×2 matrix with two absorbing states, one for each choice alternative. $\mathbf{Q}$ is an $(m - 2) \times (m - 2)$ matrix that contains the transition probabilities $p_{i,j}$ (see Equation C2). $\mathbf{R}$ is an $(m - 2) \times 2$ matrix that contains the transition probabilities from the transient states to the absorbing states.

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

$$[P(R_A|S_A),\ P(R_B|S_A)] = \mathbf{Z} \cdot (\mathbf{I} - \mathbf{Q})^{-1} \cdot \mathbf{R}, \quad \text{(C4)}$$

where $\mathbf{Z}$ is an $m - 2$ vector denoting the initial starting position of the process. Assuming no bias, then $\mathbf{Z}_{(m-3)/2+1} = 1$ with all other entries set to 0. $\mathbf{I}$ is the identity matrix with the same size as $\mathbf{Q}$, $(m - 2) \times (m - 2)$. The probability distribution function for the first passage time reaching the boundary with $t = n\rho$ $(n = 1, 2, \ldots, \infty)$ is

$$[\Pr(T = t|R_A,S_A),\ \Pr(T = t|R_B,S_A)] = \mathbf{Z} \cdot \mathbf{Q}^{n-1} \cdot \mathbf{R}./[\mathbf{Z} \cdot (\mathbf{I} - \mathbf{Q})^{-1} \cdot \mathbf{R}]. \quad \text{(C5)}$$

The CDF is

$$[\Pr(T \leq t|R_A,S_A),\ \Pr(T \leq t|R_B,S_A)] = \mathbf{Z} \cdot \mathbf{Q}^{-1} \cdot (\mathbf{I} - \mathbf{Q})^{n} \cdot \mathbf{R}./[\mathbf{Z} \cdot (\mathbf{I} - \mathbf{Q})^{-1} \cdot \mathbf{R}]. \quad \text{(C6)}$$

The mean decision time conditional on each choice is

$$[E(T|R_A,S_A),E(T|R_B,S_A)] = \mathbf{Z} \cdot \mathbf{Q}^{-2} \cdot \mathbf{R}./[\mathbf{Z} \cdot (\mathbf{I} - \mathbf{Q})^{-1} \cdot \mathbf{R}], \quad \text{(C7)}$$

where ./ indicates elementwise division. 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 a process that uses an optional stopping rule. This permits the model to predict not only the distribution of confidence ratings but also the distribution of interjudgment times.

In general the model assumes that markers κ_i are placed along the evidence state space representing the different confidence ratings (.50, . . . , 1.00), one for each rating. For the intermediary confidence ratings (.60, .70, .80, and .90), each time the judge passes one of these markers there is a probability w_i that the judge exits and gives the corresponding confidence rating. The evidence states representing the confidence rating of .50 and 1.00 were set equal to an absorbing boundary ($w_{.50} = w_{1.00} = 1.0$, thus the black circles shown in the lower chain in Figure 5).

(*Appendices continue*)

In what follows, we 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 –θ 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 except the 1.00 confidence rating (it is associated with the absorbing boundary). Assuming the boundary for Response Alternative A was reached, the state space of the second stage is

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

where in this case $\kappa_{.50} = -l\Delta$ and $\kappa_{1.00} = k\Delta$. The transition probability matrix is similar to that given in Equation C3 except for three changes. A new transition matrix $\mathbf{P}^{\mathbf{A}}$ is formed, that is, $[(m + 4) \times (m + 4)]$, where $m = k\Delta + l\Delta + 1$. The values in the transition matrix for $p_{i,j}$ are given by Equation C2. The transition matrix has an additional four rows and four columns to account for the possibility of giving a confidence rating (*conf* = .60, .70, .80, and .90) associated with one of the four confidence markers κ_{conf}. $\mathbf{P}^{\mathbf{A}}$ in its canonical form is

$$\mathbf{P}^{\mathbf{A}} = \left|\begin{array}{c|c} \mathbf{P}_1^{\mathbf{A}} & \mathbf{0} \\ \hline \mathbf{R}^{\mathbf{A}} & \mathbf{Q}^{\mathbf{A}} \end{array}\right|$$

$$= \left|\begin{array}{c|cccccccccccccc}
1 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 & 0 \\
m+1 & 0 & 1 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & & \vdots & & \vdots & & \vdots & \vdots \\
m+4 & 0 & 0 & \cdots & 1 & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 & 0 \\
m & 0 & 0 & \cdots & 0 & 1 & 0 & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 & 0 \\
\hline
2 & p_{2,1} & 0 & \cdots & 0 & 0 & p_{2,2} & p_{2,3} & \cdots & 0 & \cdots & 0 & \cdots & 0 & 0 \\
3 & 0 & 0 & \cdots & 0 & 0 & p_{3,2} & p_{3,3} & \cdots & 0 & \cdots & 0 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & & \vdots & & \vdots & & \vdots & \vdots \\
m^*_{.60} & 0 & w_{.60} & \cdots & 0 & 0 & 0 & 0 & \cdots & (1-w_{.60})p_{m^*_{.60},m^*_{.60}} & \cdots & 0 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & & \vdots & & \vdots & & \vdots & \vdots \\
m^*_{.90} & 0 & 0 & \cdots & w_{.90} & 0 & 0 & 0 & \cdots & 0 & \cdots & (1-w_{.90})p_{m^*_{.90},m^*_{.90}} & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & & \vdots & & \vdots & & \vdots & \vdots \\
m-2 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 & \cdots & p_{m-2,m-2} & p_{m-2,m-1} \\
m-1 & 0 & 0 & \cdots & 0 & p_{m-1,m} & 0 & 0 & \cdots & 0 & \cdots & 0 & \cdots & p_{m-1,m-2} & p_{m-1,m-1}
\end{array}\right|.$$

In general it takes a similar form as the probability transition matrix $\mathbf{P}$ used during the choice stage. The changes are as follows. To allow for the possibility of exiting the evidence accumulation process and giving a confidence rating, the evidence state at the row corresponding to the index of $\mathbf{P}^{\mathbf{A}}$ associated with the confidence marker c_{conf} is multiplied by $(1 - w_{conf})$. The last rows of $\mathbf{P}^{\mathbf{A}}$ contain all zeroes except for the new absorbing states associated with each confidence rating, which are set to 1 (see $\mathbf{P}_1^{\mathbf{A}}$). The last four columns of $\mathbf{P}^{\mathbf{A}}$ contain all zeros except for the row corresponding to the absorption state for each confidence rating and the row corresponding to the confidence marker, m_{conf}*, which is set to w_{conf} (see $\mathbf{R}^{\mathbf{A}}$).

Using the submatrices in the transition matrix $\mathbf{P}^{\mathbf{A}}$, the distribution of confidence ratings is

$$[\Pr(.50|R_A,S_A), \Pr(.60|R_A,S_A), \ldots, \Pr(1.00|R_A,S_A)] = \mathbf{Z}^{\mathbf{A}} \cdot (\mathbf{I}^{\mathbf{A}} - \mathbf{Q}^{\mathbf{A}})^{-1} \cdot \mathbf{R}^{\mathbf{A}}, \quad \text{(C8)}$$

where $\mathbf{Z}^{\mathbf{A}}$ is an $m - 1$ vector denoting the initial starting position of the process, the location of the choice threshold from the choice stage. The functions for expected interjudgment times can be found in the same way by adapting Equations C4 to C8.

We modified the QMP method to include the distribution of interjudgment times. To do so we calculated the .1, .3, .5, .7, and .9 quantiles for the interjudgment times for correct and incorrect confidence ratings. Because the study was not designed to investigate the distribution of interjudgment times, in some cases we have a small number of observations per confidence judgment. Therefore, we then found the average quantiles across the six subjects. The same was done for the correct and incorrect decision times. We fit the model using the predicted distributions of decision times for correct and incorrect choices averaged across the distribution of confidence ratings and the distribution of interjudgment times for each confidence rating associated with a correct and incorrect choice averaged across decision times.

We fit the full 2DSD optional stopping model to each time pressure condition of each decision task, collapsing across the different levels of difficulty. This was done for two reasons. First, in order to have enough data to fit the model at the level of the distribution of interjudgment times we had to collapse across the levels of difficulty. Second, collapsing across difficulty levels also sped up the fitting process, as fitting the Markov approximations can sometimes be fairly intensive in terms of memory needs. We fit the full model to each condition, instead of a constrained model where parameters were set equal across conditions, because preliminary fits showed that the fitting routine had difficulty handling these constraints. Nevertheless the parameter estimates in Table 9 suggest they can be set equal across conditions. Each condition had approximately 81 free data points, and the 2DSD optional stopping model had 14 free parameters.

Received May 14, 2009
Revision received January 22, 2010
Accepted January 26, 2010 ■