Psychological Review
2011, Vol. 118, No. 4, 583–613

© 2011 American Psychological Association
0033-295X/11/$12.00 DOI: 10.1037/a0025191

A Dynamic Stimulus-Driven Model of Signal Detection

Brandon M. Turner and Trisha Van Zandt
The Ohio State University

Scott Brown
University of Newcastle

Signal detection theory forms the core of many current models of cognition, including memory, choice, and categorization. However, the classic signal detection model presumes the a priori existence of fixed stimulus representations—usually Gaussian distributions—even when the observer has no experience with the task. Furthermore, the classic signal detection model requires the observer to place a response criterion along the axis of stimulus strength, and without theoretical elaboration, this criterion is fixed and independent of the observer's experience. We present a dynamic, adaptive model that addresses these 2 long-standing issues. Our model describes how the stimulus representation can develop from a rough subjective prior and thereby explains changes in signal detection performance over time. The model structure also provides a basis for the signal detection decision that does not require the placement of a criterion along the axis of stimulus strength. We present simulations of the model to examine its behavior and several experiments that provide data to test the model. We also fit the model to recognition memory data and discuss the role that feedback plays in establishing stimulus representations.

Keywords: signal detection theory, recognition memory, cognitive modeling, dynamic models of information processing

Signal detection theory (SDT) is crucial to many important theories in cognitive psychology, especially those theories that deal with performance in two-choice tasks. In such tasks, an observer is presented with a series of trials in which he or she must respond to a stimulus. The stimulus is one of two types, either "noise" (requiring a response of "no") or "signal" (requiring a response of "yes"). What constitutes noise or signal can be very flexible.

The SDT framework assumes that the presentation of a stimulus gives rise to a perception of some sensory effect in the cognitive apparatus of the observer. The magnitude of the effect, conceived on some relevant experiential scale such as "loudness" or "familiarity," is used as the basis of the "yes" or "no" decision. Random noise in the observer's perceptual system (or in the stimulus itself) results in varying magnitudes of effects over different stimulus presentations but, on average, signals result in larger effects than noise. Variability in sensory effects is represented by two random variables that often are assumed to follow equal-variance Gaussian distributions, though this assumption is not strictly necessary. To make a decision, traditional SDT assumes that observers place a criterion along the axis of sensory effect. The "yes" or "no" response is determined by whether the perceived effect is above or below this criterion (see Macmillan & Creelman, 2005, for a review).

SDT is not confined to the relatively simple problem of detecting the presence of signals. Any two-choice task that can be recast as a magnitude judgment can be shoehorned into the SDT framework. Consider, for example, a lab technician whose job is to examine Pap smears and decide which are normal and which show signs of disease. The technician looks at a number of features, such as the number of white blood cells, lymphocytes, squamous cells, presence of bacteria, and so forth. Some of these features may be more diagnostic than others of disease, but collapsing these features onto a single dimension (which we might label "abnormality") permits us to apply the SDT machinery to the problem (Beck & Shultz, 1986; Metz, Herman, & Shen, 1998).

There are two distinct ways in which SDT is used in psychology. The first use is as an analytic tool to measure discriminability (usually estimated by a statistic like d') and response bias (usually estimated by a statistic like β). The second use is as a psychological model for how people structure discrimination problems and make simple choices. When SDT is used as a model, the two distributions of sensory effect serve as the mental representations of the different stimulus classes. The decision rule is the selection of a criterion (β) separating noise from signals at some satisfactory location within the representation. Although the analytic contribution of SDT to quantifying discrimination performance cannot be disputed, there are several long-standing theoretical problems associated with the use of SDT as a model of choice that,

This article was published Online First September 5, 2011.

Brandon M. Turner and Trisha Van Zandt, Department of Psychology, The Ohio State University; Scott Brown, School of Psychology, University of Newcastle, Callaghan, Australia.

This work was made possible by National Science Foundation Grants SES-0437251, BCS-0738059, and SES-1024709 and by Australian Research Council Project DP0878858. Portions of this work were presented at the 42nd Annual Meeting of the Society for Mathematical Society, Amsterdam; the 2010 annual meeting of the Psychonomics Society, St. Louis; and the 2011 Annual Context and Episodic Memory Symposium, Philadelphia. Portions of this work were submitted by Brandon M. Turner in partial fulfillment of the requirements for the master's degree in psychology at The Ohio State University. We would like to thank Rob Nosofsky for helpful comments on experimental design and modeling and John Dunn and Michael D. Lee for helpful comments that improved an earlier version of this article.

Correspondence concerning this article should be addressed to Trisha Van Zandt, Department of Psychology, The Ohio State University, Columbus, OH 43202. E-mail: van-zandt.2@osu.edu

though acknowledged, have been generally disregarded (see also Balakrishnan, 1998, 1999) when modeling the processes that give rise to the representation or the choices based on the representation.

Since SDT's entrée into psychology, we have known that there are effects in two-choice discrimination data that it does not explain (see, e.g., Green & Swets, 1966). These include sequential dependencies over repeated responses, changes in the shape of the receiver operating characteristic (ROC) with changes in discriminability, payoffs and prior probabilities, and improvements in discrimination performance with experience. Modifications to the SDT model to explain these effects have focused primarily on how observers move the criterion with experience (Erev, 1998; Kubovy & Healy, 1977; Mueller & Weidemann, 2008; Treisman & Williams, 1984). This approach, in general, avoids the problem of how the observer acquires a representation of the statistical properties of the stimuli. Even in the modified versions of SDT, the "yes" or "no" decision is always determined by whether the perceived stimulus is greater or less than the changing criterion while the representation is assumed to be fixed.

This approach has drawbacks. First, it requires the stimulus distributions to have a monotonic likelihood ratio (the likelihood of a "signal" stimulus giving rise to the observation increases with the magnitude of the observation) so that the model only needs to explain the placement of a single criterion (Kubovy & Healy, 1977). Second, many of the solutions to the criterion placement problem have used a referent, mean, or some other fixed point around which the criterion varies (e.g., Brown & Steyvers, 2005; Mueller & Weidemann, 2008; Treisman & Williams, 1984). In laboratory situations that closely resemble real-world problems, such as recognition memory, it may be reasonable to assume that subjects have acquired accurate representations of, say, "old" and "new" (or recent and not-so-recent) objects that they can bring to bear on the laboratory task, and so also assume that subjects can place a single appropriate criterion that can then be adjusted to meet the demands of the task.

However, many laboratory tasks ask the subjects to make arbitrary classifications of unfamiliar stimuli that arise from poorly defined categories. This occurs in psychophysical discrimination tasks and some categorization tasks. If it is agreed that observers cannot possibly have a useful representation of these stimuli before significant experience with the stimuli themselves, then it must also be agreed that there is no basis for the observers to place a criterion, referent or bound within whatever impoverished representation they may possess. It seems reasonable that subjects would need to estimate the likelihood that a stimulus comes from one class versus the other, placing a criterion at the point along the decision axis where the likelihood of signals exceeds the likelihood of noise. Proposed mechanisms that depend on the likelihood ratio lack an explanation of how the perceived likelihoods arise (e.g., Erev, 1998). Most elegant theories of criterion placement, then, have simply shifted the criterion placement problem to some other component of the model.

The purpose of our article is to present an alternative approach to discrimination performance. We begin with a brief review of SDT models and the areas of psychology where SDT plays a critical role, and also of some of the more realistic SDT models of choice. We then propose a new model of two-choice decisions, rooted in the highly successful SDT framework, that may serve as a potential explanation of two-choice (detection) behavior. Our model differs from many others in that we attempt to model the dynamic characteristics of discrimination performance, changes in the decision process that arise because of changes in the stimulus environment. Our perspective is that many effects on performance arise from the evolution of stimulus representations and not, as many other models posit, because of changes in decision criteria.

Role Played by SDT in Modern Cognitive Models

It is difficult to overestimate the contribution of SDT to models of cognitive processing. These models encompass performance in psychophysics, memory, judgment and choice, and categorization, to name but a few. In this section we provide an overview of some of these models, with the goal of describing how SDT, though providing an excellent framework for the statistical characterization of performance, does not necessarily provide an explanation of how processing unfolds.

SDT spread into cognitive psychology from psychophysical research in the middle of the last century, where it proved indispensable in quantifying human performance in detection tasks. In psychophysical detection tasks, experimenters ask observers to respond to stimuli with intensities very close to perceptual limits. Many psychophysical techniques (such as the method of limits) ask people to respond when they detect the stimulus or a change in the stimulus with gradual, experimenter-controlled changes in stimulus intensity. Such techniques cannot determine whether changes in response frequency are due to the experimenter's changes or to changes in response bias.

Response bias is considerable in such tasks (Fernberger, 1920; Verplanck, Collier, & Cotton, 1952; Verplanck & Cotton, 1955). SDT, in its earliest applications to psychophysics, was very important for the reason that it provided researchers with a framework for separating bias from discriminability (e.g., Swets, Tanner, & Birdsall, 1961). This framework was so powerful and so reliable that SDT became the basis for many models of discrimination across cognitive psychology despite its lack of explanatory power (see below).

Memory

In 1958 Egan recognized that SDT could be used to quantify recognition memory performance: the distinction between old, studied items and new, unstudied items could be recast as a discrimination problem on a strength or familiarity axis instead of a stimulus intensity axis. His work applying SDT to recognition led to the class of recognition memory models based in strength theory. In these theories the SDT framework was presumed to underlie the memory decision, which was made on the basis of the perceived strength or familiarity of a target stimulus (Bernbach, 1967, 1971; Murdock, 1965; Murdock & Dufty, 1972; Wickelgren & Norman, 1966).

Later global memory models used strength theory as the decision-making mechanism in recall and recognition (Gillund & Shiffrin, 1984; Hintzman, 1988; Murdock, 1982). These models, following some considerable evolution, survive as our best explanations of these memory processes, and are some of the only models in psychology to describe how "memory strength" might arise (Dennis & Humphreys, 2001; Glanzer, Adams, Iverson, &

Kim, 1993; Shiffrin & Steyvers, 1997). Gillund and Shiffrin's (1984) search of associate memory model, for example, proposed that memory strength comes from the summed strength of match between a stimulus and the elements stored in memory, including the context of those elements. The decision of whether a stimulus is old or new is determined by the strength of the perceived match: Old items give rise to stronger matches than new items. The observer must determine a criterion level of strength to discriminate old from new items, and respond "old" to all items yielding a perceived strength greater than that criterion.

Ratcliff (1978) took the SDT representation of the strength models as a front end for his work on the memory decision process, which he characterized as a Wiener diffusion of evidence over time. Strength, in this formulation, became the rate at which the decision process drifted toward one response or the other. Many modern memory models now assume the existence of such a process after the strength computation, either implicitly, explicitly, or by using some approximately similar information accumulation mechanism (Brown, Marley, Donkin, & Heathcote, 2008; Shiffrin & Steyvers, 1997). The information accumulation mechanism does not address, however, how the strength representations are constructed over time, nor how subjects come to learn which strengths are ambiguous and could indicate an old or new item equally well. Learning which strengths are ambiguous is equivalent to determining a criterion, and this is important both theoretically and computationally. In the diffusion model, ambiguous strength values map to drift rates of zero—the problem of how to set the SDT criterion is simply shifted to the problem of setting a "drift criterion."

Current controversy over two-process memory models (Wixted, 2007; Wixted & Stretch, 2004; Yonelinas, 1994, 1999), which assume that a fast familiarity process and a slower recollective process both contribute to recognition judgments, is also firmly rooted in the tenets of SDT. Theories about the relative contribution of the two processes are based on the shape of the ROC curve estimated from recognition data. Models derived from these theories have not modified the basic ideas of SDT-as-model: Memory decisions are based on representations of strength and placement of criteria (although some theories have expanded into multiple dimensions; e.g., Cohen, Rotello, & Macmillan, 2008; Rotello, Macmillan, & Reeder, 2004).

Although memory models explain to some extent what the decision axis is in recognition memory, and how variance in memory strength arises across sets of old and new items, they rarely address the question of how a stimulus representation is built with increasing exposure to different old and new items. The representation is assumed to exist given the statistical properties of the stimulus and the existing associations between stimuli and items stored in memory, but there is no reasonable way the observer would be aware of such statistical properties without experience. Recognition memory tasks, however, can be argued to be very similar to the kinds of memory problems subjects solve in everyday life, so the assumption of preexisting, stable representations of familiar and unfamiliar things is not an unreasonable one.

One critical difference between recognition memory tasks and the other kinds of tasks to which SDT has been applied is the use of feedback during performance. Recognition memory tasks do not usually tell the subject after each response whether the response was correct, yet subjects still perform the task with high accuracy. Discrimination tasks with arbitrarily defined stimulus classes must often provide such feedback, or subjects can perform quite poorly. However, as we show in this article, subjects can respond at least appropriately, if not optimally, to changes in the stimulus classes over time. The fact that people can perform well in some tasks without feedback and only marginally well in others emphasizes the importance of a theory for how people acquire representations.

Categorization

Categorization tasks ask observers to make absolute identifications of exemplars drawn from different categories. Such exemplars may be pictures of everyday objects (such as birds, chairs, or dogs), or they may be more artificial, constructed either by making a limited number of changes to an object defined on a finite number of features (such as circular arcs of varying radii with radial lines of varying angle; see Nosofsky, 1985) or by selecting features at random from joint distributions of these features (e.g., horizontal and vertical line segments of different lengths joined as right angles; see Ashby & Gott, 1988).

The issue of whether a stimulus represents a signal or noise is, fundamentally, a simplified categorization problem. Categorization tasks, therefore, differ from traditional psychophysical tasks primarily in terms of the complexity of the stimuli. For instance, consider the task developed by Ashby and Gott (1988), in which stimulus categories were defined by the lengths of horizontal and vertical line segments joined to make a right angle (e.g., ┌). Categories were defined by the mean lengths of the horizontal and vertical lines: A-type stimuli had mean lengths of (400, 500) pixels, and B-type stimuli had mean lengths of (500, 400) pixels. Each exemplar was constructed by drawing a horizontal and vertical length from the appropriate bivariate logistic distribution.

Ashby and Gott (1988) called this method of stimulus construction the *general recognition randomization technique.* The bivariate category structure is ill-defined, as in SDT, because exemplars from the two categories may overlap considerably, making perfect performance theoretically impossible. To determine how a particular random exemplar should be categorized is akin to a signal detection problem, but in two-dimensional space. In this case, the optimal criterion (bound) is a curve in this space dividing the representation into two regions, corresponding to the two categories. Ashby (1992) provided a review of a number of different models of this type, which are usually referred to as decision bound models (DBMs; Ashby & Townsend, 1986).

Ashby and colleagues' (Ashby, 1992; Ashby & Maddox, 1993) theory assumes that the decision boundaries are random, which accounts for the fact that people usually do not seem to use optimal boundaries. In particular, the decision bound used on any trial is equal to the optimal bound plus noise. Maddox has extended decision-bound theory to explain how people learn to adjust decision bounds in response to changes in stimulus base rates and payoffs (see, e.g., Maddox, 2002, for a review). These modeling efforts have now extended into broader attempts that link categorization performance to neurological mechanisms (e.g., Ashby & Ell, 2001; Schnyer et al., 2009).

DBMs can be contrasted to exemplar-based models, such as the generalized context model (GCM), which do not use SDT-like decision bounds (Nosofsky, 1986; Nosofsky & Palmeri, 1997; Nosofsky & Stanton, 2005). Although stimuli are still represented as points in some perceptual space, categorization decisions are made on the basis of how similar a stimulus is to all the exemplars in a category. This overall similarity is computed for each category, and the probability of choosing a particular category is given by the similarity for that category divided by the summed similarities over all categories. Depending upon how similarity is computed and on how noise is added to the decision bounds, the GCM may be mathematically equivalent to some DBMs (Ashby & Maddox, 1993).

Kruschke (1992) presented a complementary approach to categorization with his ALCOVE model, which combines error-driven learning with exemplar-based representations in a connectionist network model. Similar to the GCM (Nosofsky, 1986), ALCOVE represents stimuli as points in a multidimensional space. These representations are developed in a training phase during which hidden nodes are placed at coordinates in the psychological space within which exemplars are located. Upon presentation of a test stimulus, each node is activated according to the psychological similarity of the stimulus to these hidden nodes. The psychological similarity function is the same as in the GCM—a Minkowski distance (Nosofsky, 1986). To elicit a response, the hidden nodes are connected to each possible response category, represented by category nodes, through association weights. The activation of a category node is the weighted sum of activations of the hidden nodes. The weights are learned by gradual adjustments based on error between the category judgment and feedback provided during training. Like the GCM, the probability of a category judgment is equal to the activation of that category node divided by the sum of the activations for all categories.

The new model we present in this article shares common features with the DBM, the GCM, and ALCOVE but, in its most general form, is equivalent to none of these. We discuss the relationship between our model and these other models in the General Discussion. For now, it is sufficient to state that whereas the DBMs, the GCMs, and ALCOVE focus on explanation of asymptotic performance after the construction of stable representations obtained through training, we are concerned with early changes in performance resulting from experience and changes in the stimulus stream.

Confidence

The SDT model has also contributed to modeling and empirical work in judgment, where observers are asked (either after a dichotomous response or in place of one) to provide an affective value describing the extent to which they endorse one response or the other. For example, given an item that may or may not have been presented earlier, an observer may be asked to indicate on a 50%–100% scale the probability that the item was presented. The response scales in such experiments take many forms, from the half-scale 50%–100% (usually used when a dichotomous response is also made) to the full-scale 0%–100% (allowing the observer to endorse either response option) to Likert-type scales (1 = *confident noise*, 2 = *probably noise*, 3 = *maybe noise*, . . . 6 = *confident signal*). Such responses are assumed to reflect the observer's confidence, either in his or her response or in the identity of the item.

Egan (1958) originally proposed the use of such rating scales as a way to reduce the number of trials necessary to construct an ROC curve. Assuming the strength of sensory effect could be mapped directly to judgment ratings, the different rating levels corresponded to decisions made with different criterion placements. However, despite the similarity between the ROCs constructed with rating scales and those constructed with different payoff schemes, the judgment ratings do not seem directly based on the strength of sensory experience (Van Zandt, 2000). Modern models of confidence, like modern memory models, assume the SDT front end to an information-accumulation decision process, and that the judgment is based on characteristics of the system state at the time of the decision. Vickers (1982) and later Van Zandt (2000) and Merkle and Van Zandt (2006) modeled response confidence from the balance of evidence in race-type models, where levels of evidence are stored separately for each response. More recently, Pleskac and Busemeyer (2010) and Ratcliff and Starns (2009) have presented similar modeling approaches using SDT-like assumptions to describe the strength of information.

Research on confidence calibration has also relied on SDT. Confidence calibration refers to the ability of observers to gauge the accuracy of their own responses: People tend to be overconfident in their answers to difficult decisions and underconfident in their answers to very easy decisions (Dawes, 1980; Dunning, Heath, & Sols, 2004; Lichtenstein & Fischoff, 1977). When asked, for example, to respond to a general knowledge question such as "What is the capital of Australia?" people overestimate the probability that their choice is correct. The earliest quantitative model to address confidence calibration was Ferrell and McGoey's (1980) decision variable partition model. They applied this model to two-part decisions: Their subjects first made a "yes" or "no" decision and then reported their confidence that the decision was correct. They assumed that the presentation of a stimulus gives rise to a sensation of certainty for each of the possible response options. The option with the largest subjective certainty is chosen, and the judgment of confidence is based on the difference between these subjective certainties. The distribution of the differences depends on whether the response was correct, leading to two Gaussian distributions like those of SDT. Criteria placed along the decision axis (representing subjective certainty difference) allow the subjects to map sensations of difference into levels of confidence.

Wallsten and González-Vallejo (1994) proposed a model with a similar relationship to SDT, called the stochastic judgment model. This theory explained data from paired-comparison and true or false sentence verification tasks by assuming that an internal and very accurate confidence level is perturbed by random noise. Confidence in the truth of true statements is greater than confidence in the truth of false statements, leading to two distributions of overt confidence. Decisions about whether statements are true or false are determined by the placement of a criterion along the confidence scale; judgment biases are explained by how the criterion is placed.

More recent models (including Brenner, 2003; Pleskac & Busemeyer, 2010; Ratcliff & Starns, 2009) feature, in some form or other, an SDT-like distribution of internal confidence values with cut-point criteria, and these elements are fundamental to the models. However, although SDT has played an important role in models of judgment, none of these models has addressed the fundamental problems of SDT-as-model: the formation of the stimulus representation and placement of criteria. Work by Vickers and Lee (1998) and Lee and Dry (2006) has addressed this issue by taking a different approach with their "self-regulating" accumulator model, which dynamically monitors decision confidence and occasionally adjusts decision thresholds with the aim of maintaining a target level of confidence. We discuss this model in more detail below.

SDT Models With Dynamic Criteria

As we have described, there are two conceptual problems in using the traditional SDT framework as a model of choice. First, we must assume that an accurate stimulus representation exists for each observer, even when the observers have had no prior experience with the task, and that these representations do not change with the observers' experiences. In other words, the parameter d' representing the discriminability of the stimulus is fixed. The second problem, which is related to the first, is that we must assume that observers can accurately place a criterion within the representation and that this criterion (β) is either fixed for the duration of the experiment or adjusted with mechanisms that depend on existing representations (e.g., Erev, 1998; Maddox, 2002; Mueller & Weidemann, 2008; Treisman & Williams, 1984).

One implication of representations and parameters that are fixed over trials is that decisions made by the observer are independent of previous decisions and their outcomes. However, there is empirical evidence, extending back over many decades, to demonstrate that this is not true. For example, Howarth and Bulmer (1956) asked observers to perform a detection task in which an auditory signal was presented in white noise. Treisman and Williams (1984) later analyzed Howarth and Bulmer's data and found strong positive sequential dependencies such that observers tended to repeat responses they had made on earlier trials. Treisman and Williams then found similar effects in their own data. Their additive learning model was an attempt to explain these effects.

Extensions to the SDT model have focused almost exclusively on the criterion problem. Very early work used stochastic learning mechanisms to describe how response probabilities changed with feedback and reinforcement (e.g., Atkinson & Kinchla, 1965; Dorfman & Biderman, 1971; Kac, 1962). Whether criteria are assumed to change systematically with the stimulus environment (e.g., Brown & Steyvers, 2005; Kubovy & Healy, 1977; Treisman & Williams, 1984) or randomly (e.g., Benjamin, Diaz, & Wee, 2009; Dorfman, Saslow, & Simpson, 1975) depends on the task to which the model is applied. Other models have been designed to explain changes in performance arising from both changes in the stimulus representation and criterion adjustments (e.g., Kubovy & Healy, 1977; Lee & Dry, 2006). We address each of these models in turn.

Additive Learning Models

Persistent sequential dependencies led to the development of several dynamic models to explain performance in signal detection and categorization tasks (Brown & Steyvers, 2005; Kac, 1962, 1969; Petrov & Anderson, 2005; Rabbitt, 1981; Treisman & Williams, 1984; Vickers & Lee, 1998, 2000). These models explain sequential dependencies in a number of ways. For example, Kac (1962, 1969) proposed an error-correcting model in which an observer's criterion could be altered from trial to trial but only after an error. If the observer's response was "yes" when noise was presented, the criterion shifted upward to prevent this error in subsequent trials; if the observer's response was "no" when a signal was presented, the criterion shifted downward. Similar models were proposed by Dorfman and Biderman (1971) and Thomas (1973). Dorfman et al. (1975) later expanded Kac's model to allow for random fluctuations in the criterion from trial to trial.

These additive learning models perform error-rate monitoring: The criterion moves to minimize error. Treisman and Williams (1984) explored this sort of mechanism, which requires the observer to be presented with feedback, in some detail. They focused on three models: a linear additive learning model, an exponential additive learning model, and an independent trace model. The linear and exponential additive learning models shift the criterion relative to some "reference" criterion depending on the response on each trial. The criterion shifts so that the frequency of the response just made will increase or decrease, resulting in positive or negative dependencies in the response sequence. The independent trace model is slightly different but follows the same structure. Each response results in a "memory trace" for the criterion shift (positive or negative, depending on the response). This trace decays toward zero with time. The criterion location at any point in time is a function of the summed memory traces. Like the linear and exponential additive learning models, this model produces positive or negative dependencies in the response sequence, depending on the values of the model parameters.

Change Detection

Another stimulus-based adjustment mechanism was used by Brown and Steyvers (2005) to model change detection in discrimination tasks. They examined performance in a lexical decision task and a numerosity judgment task—two tasks typically treated within the SDT framework. Subjects performed one of these tasks with stimuli that were either easy or hard to discriminate. Within blocks of trials the task environment shifted unpredictably from easy to hard, or vice versa.

Brown and Steyvers (2005) proposed a model of dynamic criterion adjustments in which two a priori stimulus representations and decision criteria (one set for the easy environment and one set for the hard environment) determine decision-making performance. The model assumes that after switching from one environment to the other, the subject continues to use the criterion he or she was using before, until the subject recognizes that the task has changed. The critical parameter is the lag, the time it takes the subject to switch from one criterion to the other.

Random Criteria

Many studies have found that systematic changes in criteria with experience are not sufficient to account for the variability in the data (Dorfman & Biderman, 1971; Dorfman et al., 1975; Kac, 1969; Kubovy & Healy, 1977; Schoeffler, 1965). The criterion must also be random. Such variability has been modeled in a number of ways, most commonly by adding noise to fixed criteria (see Ashby, 1992; Ashby & Maddox, 1993).

Erev (1998) was concerned about how payoffs influence the movement of the criterion. He proposed a cutoff reinforcement learning model that incorporated both systematic and random shifts of the criterion. Under the assumption again that observers have some basis for computing a likelihood ratio, the model postulates a set of prior "propensities" to select one criterion along a continuum. The probability that one criterion is selected is computed with Luce's (1959) choice rule: the propensity for that criterion divided by the sum of the propensities for all criteria. Following a response and any payoff received, the propensities are updated in such a way that the criterion selected for the response is generalized (increasing the propensities to neighboring criteria) and reinforced (increased to the extent determined by the payoff).

This scheme results in a probability distribution over criterion locations that shifts with experience and payoffs. Erev (1998) examined a number of published data sets and showed that it provided a good explanation for their results, including the change in the proportion of apparent static criterion violations (e.g., responding "yes" when the stimulus was apparently below the criterion), changes in d' over time, sequential effects, and the extent of deviations from optimality. However, the model is, like the ideal-learner model, restricted to choice situations where a single criterion is sufficient for a response. That is, the likelihood of a signal is monotonically increasing with stimulus magnitude. How the perceived likelihood values arise is also not addressed by the model.

Mueller and Weidemann (2008) developed a random criterion model that focused on the confidence rating procedure by which ROC curves are commonly generated. These ratings are assumed to require the placement of several criteria. Mueller and Weidemann's goal was to show that changes in the shape of the ROC curve with bias could be explained by criterion variability. In addition, they argued that Balakrishnan's (1998, 1999) findings that criteria do not seem to move with changes in bias were also due to criterion variability.

Mueller and Weidemann (2008) represented criterion noise with Gaussian distributions placed along a decision axis. There is a central "classification" criterion that determines whether a stimulus will be given a "yes" or "no" response (i.e., which side of the confidence scale will be used). This classification criterion is selected at random from a Gaussian distribution. Another criterion from the lowest confidence category in the appropriate direction (to the left of the classification criterion if the perceived magnitude was less than that criterion and to the right otherwise) is then sampled and compared to the perceived magnitude. If the criterion is of greater absolute magnitude than the perceived magnitude, then sampling stops—the stimulus is within a category class. Otherwise, sampling continues with the next highest confidence category.

Mueller and Weidemann (2008) fit the decision noise model to new and previously published data and showed that it could explain changes in the shape of the ROC curve as well as changes in Balakrishnan's (1998) statistics to characterize discriminability and bias. However, like the previous models we have discussed, it provides no explanation for how stimulus representations are constructed. Furthermore, although it provides a structure for criterion variability, it does not address the criterion placement problem. Where criteria are placed is determined by the means of the distributions from which criteria are sampled, but how these means are selected in the first place cannot be explained by this model.

Another variable criterion model is the noisy criterion theory of signal detection (NC-TSD) described by Benjamin et al. (2009). In their model, Benjamin et al. assumed that an observer's criterion on a particular trial is random. In particular, for k-level confidence judgments in recognition memory, $k - 1$ criteria are sampled from a Gaussian distribution, and confidence is determined in the usual way (e.g., Egan, 1958).

Similar to that of Mueller and Weidemann (2008), the NC-TSD model provides an explanation for changes in the ROC curve under bias manipulations (Balakrishnan, 1998; Van Zandt, 2000). Like Mueller and Weidemann, Benjamin et al. (2009) explained this result with criterion variability, but Benjamin et al. assumed that this variability is a function of bias. Specifically, as the criterion shifts farther from the location of the unbiased criterion, the variability of the distribution from which the criterion is sampled increases.

Assuming that both stimulus and criterion variability are present makes estimation of the model parameters more difficult. To disentangle these two sources of variability, Benjamin et al. (2009) used an ensemble recognition task (similar to Nosofsky, 1983) in which the stimuli on each trial consisted of a random number of words of the same type (i.e., one, two, or four words, all of which were old or new). The size of the ensemble should not affect the criterion placement or variability. However, because stimulus variability decreases as the number of words increases, the number of words in the ensemble should affect discriminability. Benjamin et al. tested this assumption by comparing a zero criterion-variability model with a more general nonzero criterion-variability model. They claimed that the superior fit of the nonzero criterion-variability model provides evidence that the stimulus and criterion variabilities are independent.

Benjamin et al. (2009) showed that changes in the ROC with bias, such as those demonstrated by Balakrishnan (1998) and Van Zandt (2000), can be explained by criterion variability. These results also suggest that criterion noise is important for explaining the range of effects observed in recognition memory data (see Glanzer et al., 1993, for a review). Nevertheless, the NC-TSD model does not have a mechanism to explain how the distributions of evidence are established or evolve as the experiment progresses. Furthermore, similar to that of Mueller and Weidemann (2008), the NC-TSD model does not completely address the problem of criterion placement, because the means of the distributions from which criteria are selected must be fixed in advance.

The Ideal-Learner Model

Kubovy and Healy (1977) proposed an alternative to models attempting to account for behavioral dynamics with change only in the criterion, one they called the ideal-learner model. This mechanism, which is similar in some ways to our own, partially solves the problem of where stimulus representations come from by assuming that observers maintain estimates of the means of the signal and noise distributions, estimates computed from the feedback obtained after every stimulus presentation. The criterion is placed at the midpoint between the two means, and so it shifts as the means shift. The means serve as the stimulus representations, capturing the central tendency of the stimulus stream but no more detailed information.

The ideal-learner model is only useful when the signal and noise distributions have a monotonic likelihood ratio—when the ratio of signal to noise likelihoods increases with increasing stimulus strength. It cannot, for example, be applied to problems where the variance of one distribution is very different from the other, creating situations that require the placement of two criteria (e.g., our Experiment 3). Similarly, because it only stores the means and always places the criterion at the midpoint between them, it cannot explain changes in criterion placement with changes in prior probability or payoffs. Kubovy and Healy (1977) also determined that the changes in the criterion from changes in the mean estimates were not enough to explain their data: The criterion also had to vary randomly.

The Self-Regulating Accumulator

In contrast to criterion adjustments based on error, other models make criterion adjustments based on decision confidence. Most notably, Vickers and Lee's (1998, 2000) self-regulating accumulator assumes evidence accumulation mechanisms not just for the underlying decision itself but also for the metacognitions that take place across a series of trials related to overall decision confidence (see also Vickers, 1979). These metalevel accumulators monitor the confidence with which decisions are made and keep track of differences between a target confidence level and the observed confidence level. Across a series of trials, the metalevel accumulators make occasional adjustments to the primary decision thresholds, between trials, to bring confidence toward its target.

On a single trial, decision making occurs as in any other information accumulation model: The evidence accumulators for the decision gather data from the SDT front end, and the first accumulator to reach threshold determines the decision and the response time. However, after each decision, confidence in the response is also calculated, based on the levels of information accumulated toward each response. The model computes the difference between this confidence and the target confidence level, and updates the metalevel accumulators using this difference. In this way the metalevel accumulators keep track—across many trials—of over- and underconfidence. As with the decision accumulators, a response is triggered whenever a metalevel accumulator exceeds a threshold. In the case of the metalevel accumulators, however, the "response" is to adjust the decision threshold for the primary decision accumulators, thereby adjusting the confidence of decisions.

This model has some similarities to our work, in that the criteria of the decision mechanism are adjusted through experience with the stimulus stream. Among other things, Vickers and Lee's (1998) model makes an attempt like ours: to describe how a decision maker might adapt to changes in the environment, including those changes that occur in the initial stages of an experiment.

Vickers and Lee's (1998) model also differs from ours in important ways. The self-regulating accumulator model characterizes a more complex cognitive process than ours, but it also goes beyond our work by making predictions for response times and decision confidence. The self-regulating accumulator model also differs from our model in the way information about the environment is gathered. The central goal of our model is to build the internal distributions associated with the different stimulus classes. Once built, the subjects might attempt to use the representations to perform a number of tasks; we have described making simple discrimination decisions, but the representations might just as well be used for other purposes. However, Vickers and Lee's central goal was to estimate the effect that these distributions have, via decision processes, on decision confidence. Disambiguating these two approaches is an interesting problem for future research, which could, for example, be approached by investigating whether decision makers have access to any information about the evidence distributions that could not be gained from knowledge of decision confidence alone.

The self-regulating accumulator model is important because, unlike the other models we have presented in this section, it provides a unified account of stimulus learning and adaptation of criteria. There are other approaches that we could discuss, such as Maddox's (2002) criterion learning, that focus on the factors that influence how decision boundaries change with experience, payoffs, base rates, and so forth. All these models represent an advance over static SDT, which assumes that the criterion might change between experimental conditions but makes no attempt to describe how, or when, these changes might occur. Unfortunately, as Mueller and Weidemann (2008) pointed out, these models have not been implemented in cognitive models such as the ones we described above, possibly because they add additional theoretical complexity at a locus in processing that is not of immediate interest.

Many of the dynamic criterion models have not solved the two problems that are the focus of this article: Either the models assume that observers bring accurate stimulus representations to the task, representations that are fixed after training and do not change with further experience, or the criterion placement problem is not solved but only moved to another model component. The ideal-learner and self-regulating accumulator models, which do not suffer from these problems, are applicable to restricted or unique discrimination tasks (e.g., those with monotonic likelihood, equal payoffs, and equal stimulus probabilities, or those in which confidence is measured).

As with the ideal-learner and self-regulating accumulator, we considered the problems of stimulus representation development and dynamic criteria simultaneously. However, we were particularly interested in explaining changes in discrimination performance with changes in the stimuli over trials. This resulted in a new dynamic SDT model that describes how signal and noise representations are constructed and evolve over time. This model does not require the observer to set an explicit response criterion within the representation (although it does require the observer to make an evaluation of relative likelihood). In what follows, we present the new model and then discuss the results of both simulation and empirical studies to test the model. For the simulation studies, we apply SDT analyses to simulated data to show that our model behaves (via changes in d', β, and the ROC curve) like human observers under changes in discriminability and factors that influence response bias. The experiments are designed to provide a set of data to which the model can be fit, and in particular, we manipulate factors that influence discriminability and response bias over trials. This allowed us to explore the ability of the model (and the observers) to respond to sometimes drastic changes in the stimuli. We show, in several stimulus contexts, that our model provides a good explanation for changes in choice performance over time and across conditions.

The Model

Up to this point, we have discussed the role that SDT has played in the development of cognitive models and different approaches modelers of signal detection performance have taken to explain dependencies across sequences of responses. Now we introduce a new model that takes a different approach to this problem. In our model, we do not assume that subjects have an accurate stimulus representation before actually performing the task. Indeed, we do not even assume that subjects have a very rich representation of the axis underlying the decision. The subjects begin by constructing a prior that conforms to previous similar experiences and the instructions from the experimenter, and gradually build the representations on the basis of experience with the stimuli as they are presented.

There are four components of the model. First, we assume either that the subject brings a useful, informative prior from past experience to bear on the task or that the subject can establish a minimally informative prior, based on the experimenter's instructions, over an impoverished representation of the decision axis. By *impoverished* we mean that the observer can store likelihood information only for a limited number of points along the axis, although the location of those points may change over time. Second, we assume that the subject updates these priors on the basis of available information: the stimulus presented, the response to that stimulus, and feedback (if provided). The updating process relies on an adaptive kernel estimate—similar to a running average—to increase the subject's estimate of the likelihood of a signal (or noise, as appropriate) around the magnitude of the currently presented stimulus.

Third, we assume that the "yes" or "no" decision is, as in classic SDT, determined by the likelihood ratio of a presented stimulus. If the estimated likelihood that the stimulus is noise is greater than the estimated likelihood that the stimulus is a signal, then the subject should respond "no" (as in, e.g., Glanzer et al., 1993). However, because the stimulus representations are neither Gaussian nor unimodal, the likelihood ratio is not likely to be monotonic with perceived stimulus intensity. Therefore, there will usually be no single criterion placement along a single dimension that describes the response strategy.

Fourth, because the representation exists for only a few values along the decision axis, stimuli will give rise to perceptual effects for which no likelihood exists. In this case, we assume that the observer can access the closest location for which likelihoods do exist. The "yes" or "no" decision is made on the basis of the likelihood at that location.

Establishing the Priors

Consider the case where the task is completely novel and that observers have no useful prior. Before the first trial, the observer has only a poor idea of what signals and noise look like. However, on the basis of the experimenter's description of the task, the observer constructs a minimally informative prior for the representation. This representation will require both a support, which is some idea of the range of stimuli that might be experienced, and an initial likelihood for signal and noise stimuli at a few points on that support.

For example, the task that we use asks subjects to decide whether a patient is healthy or sick on the basis of the value of a single blood assay that varies from 1 to 99. Sick patients have higher blood assay values than healthy patients. Subjects use this information to set up a coarse representation of the number line between 1 and 99. For example, subjects could choose to store only a handful of representation points along this line (e.g., perhaps one subject chooses to store likelihood estimates for the values 10, 20, . . . , 90). If the experimenter says that stimuli will be chosen at random from the numbers between 1 and 99, without any additional information about how large signals will be relative to noise, it might be reasonable for the subject to assign equal and uniform prior likelihoods to each of the representation points. If the experimenter says that noise stimuli will have a mean of 40 and signal stimuli will have a mean of 60, the subject might select only two representation points at 40 and 60 and place symmetric, unimodal priors over them. For our purposes we assume, before the experimenter provides information about means, a uniform prior in which all values have an equal chance of being from the signal (*S*) or the noise (*N*) distributions. Depending on the shape of the priors (uniform, symmetric, or something else), the subject may not begin with a representation of the stimulus that permits a criterion placement in the usual sense.

Figure 1A shows how the prior representations might be established before the subject listens to the experimenter's instructions. In this example, the priors are defined at only 10 points, randomly spaced along the decision axis. Both the signal and noise priors are uniform and equal to one another, so the signal likelihoods (open circles) are plotted on top of the noise likelihoods (asterisks). Thus, initially, for any stimulus presented there would be an equal likelihood that it is either signal or noise.

A problem arises, however, in deciding how many representation points to store. Human observers are notoriously bad at keeping track of large amounts of information, and so we cannot expect an observer to maintain a dense representation of the support over time. Furthermore, in many situations, the stimulus stream may change over time, drifting across the support in response to changes in the task environment. For example, in our Experiment 4, the means of the signal and noise distributions shift in the middle of the experiment. Inflexible representation points do not allow observers to react to such changes, and therefore we

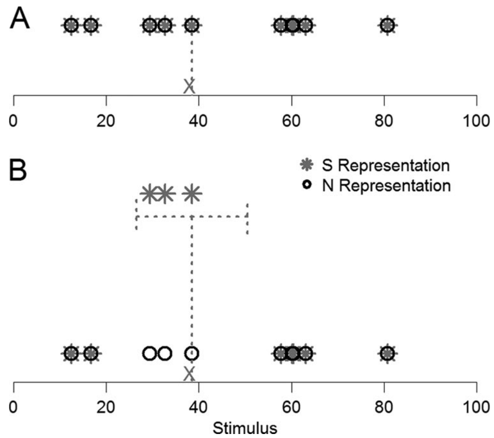

Figure 1. The prior representation of the stimuli before experience (A) and modification after the presentation of a signal (B). Signal likelihoods are plotted as open circles at every representation point, and noise likelihoods are plotted as crosses at the same points. The X is the value of the signal presented on the first trial. The likelihoods at all points within the bandwidth centered at X, shown as the horizontal interval above X, are updated following Equation 2.

wanted a mechanism where the locations of the representations could change appropriately (cf. Brown & Steyvers, 2005).

For this reason, we used representations with points that could shift. We assume that the observer starts with a small number of points (e.g., two), perhaps established by the experimenter's instructions. As stimuli are presented, new representation points are established at the values of those stimuli until a maximum number of representation points $N_{\max}$ is established. Thereafter, on every trial, with some probability γ, a representation point drops out to be replaced by a point at the value of the stimulus presented on that trial. This new point inherits the likelihoods of its closest existing representation point.

This mechanism ensures that if an observer maintains representations at $N_{\max}$ points, those $N_{\max}$ points will tend to be located at the values of the stimuli most recently presented. When a representation point is dropped, both the location of the point and the likelihood information stored about the signal and noise distributions at that point are lost. In this way the representations are free to shift as dictated by the statistical properties of the stimulus stream. This process also means that the representation points will tend to cluster where the stimuli are the most dense and hence where small changes in likelihood will be most diagnostic.

Learning the Representations

The signal and noise representations must be updated after every trial to reflect the observer's experience with the stimulus stream. The updating procedure averages the result of the new trial with the results of all previous trials. This procedure was inspired by a kernel density estimation process, which is a nonparametric approach for estimating the density of a random variable from an independent and identically distributed sample of data (Silverman, 1986). Because it is nonparametric and requires minimal distributional assumptions, it is attractive for modeling the evolution of stimulus representations that might take complicated forms.

For a sample $\{x_1, x_2, \ldots, x_n\}$, a kernel estimator has the form

$$f_n(x;\, h) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right). \tag{1}$$

The function K is the kernel, and h is a smoothing parameter or bandwidth. The kernel is usually chosen to be unimodal and symmetric about zero. Unimodal and symmetric kernels place decreasing weight on observations x_i farther from the point x at which the density is being estimated.

The kernel can take many forms; however, in this article, we will consider only the rectangular kernel,

$$K\left(\frac{x - x_i}{h}\right) = \begin{cases} \dfrac{1}{2h} & \text{for } |x - x_i| \le h \\ 0 & \text{else,} \end{cases}$$

which, we argue, has the simplest psychological interpretation. For a rectangular kernel, the density estimate at a point x is obtained by centering the kernel, a rectangle, over x and adding the contributions of each observation x_i within the rectangle. The farthest a point x_i can be from x and still contribute to the estimate at x is the bandwidth h of the kernel, or half the width of the rectangle. Large bandwidths will allow more distant observations to influence the estimate at x; smaller bandwidths will attenuate the effects of distant observations. As the bandwidth increases, the estimate will get smoother.

Equation 1 can be expressed recursively, supporting a natural interpretation as a "dynamic" kernel density estimator, where stimulus representations are updated one observation at a time. Concretely, assume a stimulus y_n is sampled from the S distribution on trial n. After the observer's response to y_n, the likelihood for one representation is increased while the other is decreased. The choice of which representation is increased will be determined by whether the observer receives feedback on the accuracy of his or her response. We discuss this in greater detail later; for now, for simplicity, we will assume that the observer receives accurate feedback and so chooses to increase the likelihood of the representation corresponding to the distribution from which the stimulus was sampled.

If the observer receives feedback that y_n was sampled from the signal distribution, then, following Equation 1, the S representation $\hat{f}_{S,n}$ at all points x will be updated by

$$\hat{f}_{S,n}(x;h) = (1 - \lambda)\hat{f}_{S,n-1}(x;h) + \lambda K\left(\frac{x - x_r}{h}\right), \tag{2}$$

where $\lambda \in [0, 1]$. At the same time, the N representation $\hat{f}_{N,n}$ decays by

$$\hat{f}_{N,n}(x;h) = (1 - \lambda)\hat{f}_{N,n-1}(x;h). \tag{3}$$

(Note that for all values of x not within the bandwidth of x_r, that is, for $x \notin [x_r - h, x_r + h]$, both representations will decay.) Similarly, if the observer was told that the sample y_n was obtained from the N distribution, then the representation of the N distribution $f_{N,n}$ would be updated and the representation of the S distribution $f_{S,n}$ would decay.[1] Note that these assumptions are equivalent to assuming that a bivariate distribution is estimated simultaneously over magnitude and stimulus identity (S or N), but that the bandwidth on the stimulus identity dimension is zero.

The parameter λ is an attention weight that determines the extent to which the updated estimate $\hat{f}_{S,n}$ (x; h) relies on the new stimulus y_n (through x_r) versus the previous estimate $\hat{f}_{S,n-1}$ (x; h). If $\lambda = 1/n$ (n, the trial, is also the number of stimuli observed), the recursive estimator becomes just a reexpression of the standard kernel density equation. We used a fixed value for λ, not depending on n, which gives the model a "memory limit" equivalent to giving exponentially decreasing weight to older items. Larger values of λ lead to steeper exponential declines, so that more weight is given to recent stimuli and less weight is given to old stimuli. In stationary environments—where the stimulus distributions do not change throughout the experiment—smaller values of λ prevent the model from paying too much attention to random fluctuations in the stimulus stream. This makes the representation less sensitive to transient changes in the stimulus environment. However, in a dynamic environment (see Experiments 2 and 4),

[1] This recursion can also be expressed as a nonlinear updating of the likelihood ratio at each point. Such an interpretation reduces the assumed memory load of the model from two numbers at each representation point to just one. For ease of exposition, we will assume that separate likelihoods for S and N are retained.

the stimulus representations should change when the stimulus environment changes, and so a higher λ can be used to allow the representations to adapt.

Using a fixed value for λ also means that the likelihoods $\hat{f}_{S,n}$ (x; h) and $\hat{f}_{N,n}$ (x; h) are not kernel density estimates at all. Instead, Equations 2 and 3 describe a simple learning rule, in which the strength of the connection between a particular location along the decision axis (x) and the response S (or N) is increased when a signal (or noise) is presented nearby. This is simply a Hebbian learning algorithm. The simultaneous decrease in the strength of all other connections means that the updating rule is similar to Oja's rule (Oja, 1982) and that we could, after some modifications, recast the model as a self-organizing neural network.

Figure 1B shows the updating process after the observation $y_1 = 38$ is identified through feedback as a signal. The likelihoods at the representation points in the S representation within the bandwidth of x_r (the closest point to y_1 is $x_r = 39$) are all increased by a factor of λ, and all other likelihood estimates (in both S and N representations) are decreased by a factor of $1 - \lambda$.

As a consequence of the restricted number of representation points and the attention weight λ, the representations do not form a true joint probability density because they do not integrate to one. With feedback, as n increases, the estimated signal (noise) likelihood at a point x is proportional to the number of signals (noise) presented within the bandwidth around x. This means that the representations are updated in the same proportion as the occurrence of signals and noise in the neighborhood of x. With accurate feedback, the lack of true joint probability densities is not a problem, because the ratio of the likelihoods at the point x will approach the signal-to-noise odds ratio. Furthermore, if λ is small, the representations will approach the distributions of the signal and noise stimuli (see Figure 2).

Without accurate feedback, the model's ability to maintain accurate likelihood ratios will be impaired. However, human observers are able to perform many discrimination tasks quite well without feedback (e.g., recognition memory and some psychophysical discrimination tasks; Hoffmann, Pizlo, Popescu, & Price, 2007; Slotnick, 2010; Van Es, Vladusich, & Cornelissen, 2007). These tasks, however, have in common the fact that observers have a lot of experience judging, say, relative levels of familiarity or stimulus intensities in real-world situations. Therefore, it is not unreasonable to assume that observers in these kinds of tasks, tasks for which the support of the stimulus representations is well defined, bring very strong priors into the laboratory. In addition, the experimenter's instructions about the nature of the stimulus distributions may have very important effects on these priors.

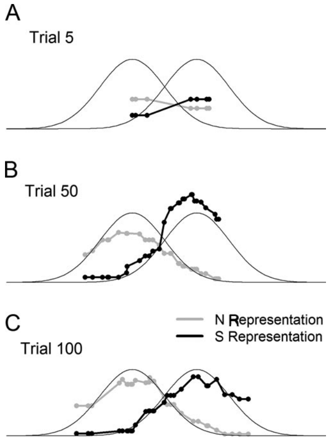

Figure 2. The evolution of the N (gray lines) and S (black lines) stimulus representations. The smooth lines show the Gaussian distributions from which stimuli were sampled. Figures 2A, 2B, and 2C show the likelihoods after 5, 50, and 100 trials, respectively.

We assume that in the absence of feedback, the signal (or noise) representation will be updated according to the response made to the stimulus; that is, the observer provides his or her own response feedback by assuming the response was correct. This means that the proportion of times the, say, signal representation is updated will reflect the number of signal responses, but not necessarily the proportion of signals presented, which may lead to a signal representation that is either overly large or too small relative to the noise representation. Given a sufficient number of signal responses, the signal representation may completely dominate the noise representation, which will decay to zero more or less rapidly, depending on the size of λ.

There are other updating rules that can prevent one representation from completely dominating the other. One such rule is for the observer to update the signal representation if the current stimulus is greater than the last stimulus presented and to update the noise representation otherwise. This strategy works quite well at preventing one representation from taking over, but tends to result in probability matching to 50% signals and 50% noise. If the signal-to-noise proportion is different from 50–50, then this updating strategy will not produce optimal performance. For purposes of this article, we will not explore this or other alternative rules, because with suitable parameter values the response feedback scheme works reasonably well and, as we show later, provides some insight into how observers respond in different task environments.

The Choice Process

Our model assumes, like classic SDT, that decisions about signals and noise are determined by the likelihoods of each for a particular perceived stimulus magnitude. However, the representation of the stimulus is constructed over time, and the observer only has estimates of the likelihoods at a few representation points. In contrast to other dynamic SDT models that focus on the location of a response criterion (Kac, 1962, 1969; Rabbitt, 1981; Treisman & Williams, 1984), we assume that the presentation of a stimulus causes the observer to access the signal and noise likelihoods at the location in the representation closest to the perceived stimulus magnitude. The observer then selects the response according to whichever likelihood is highest (and possibly by how much).

Picking the response with the highest likelihood can, with monotonic likelihood ratios, be equivalent to setting a criterion at the point where the two stimulus representations cross—where the likelihood ratio is equal to 1. Similarly, under varying payoffs, an observer may choose not to pick a response until its likelihood is much higher than that of the alternative response. With monotonic likelihood ratios, this strategy is equivalent to setting a criterion to the left or the right of the point where the representations cross. In our framework, the likelihood ratio need not (and indeed will probably not) be monotonic with stimulus strength, depending as it does on a rough and sparse estimate of the signal and noise distributions. Therefore, there may be no fixed criterion in the usual sense, and there may be more than one stimulus magnitude for which the likelihood ratio is equal to 1.

On the first trial, the observer is presented a stimulus and, on the basis of the likelihoods at the representation point closest to the perceived stimulus value, makes a decision about the distribution from which the stimulus came (*S* or *N*). If a uniform prior over the representation points is used, this first choice is a guess. Following the guess, the observer is or is not given feedback telling him or her whether the stimulus was a signal or noise. The observer then updates his or her stimulus representations according to the procedures described above. When the observer sees the next stimulus, the observer evaluates it using the updated—but still primitive—representation. Over the course of the experiment, with accurate feedback, the stimulus representations can approximate the shape of the stimulus distributions, and consequently the observer's decisions become more and more consistent (see Figure 2).

To simulate (or fit) this model, we must make a number of important choices. One is the maximum number of representation points $N_{\max}$ the observer retains (a variable that may depend on such individual differences as memory span or intelligence, or experimental factors such as attentional load). Smaller numbers of points will result in more sparse representations, which in turn will lead to more variable, less accurate choice behavior. For this article, we assume that observers maintain a small number of points in the range from 1 to 99. Observers begin with only two points, the means of the *S* and *N* distributions provided by the experimenter's instructions. As stimuli are presented, new representation points are established at the values of the stimuli until a maximum (e.g., $N_{\max} = 10$) is attained. Each new point inherits the likelihoods of the nearest established point. After the maximum number of points is reached, with every additional stimulus presented, there is some probability γ that the oldest point drops out and the new stimulus value is added as a new point. In this way, the representations are stable yet able to move with changes in the stimulus stream. Furthermore, representation points will be clustered in the regions of highest stimulus probability, yielding more diagnostic likelihood information and better discriminability between signals and noise.

To the extent that representations become smooth with extended experience with the stimulus set, it may be reasonable to assume that observers can retain likelihoods (or just likelihood ratios) at a number of points $N_{\max}$ that is somewhat greater than the usual short-term memory span of 7 ± 2. This is because not only are the representation points ordered along the decision axis, so that there is structure that relates the different representation points, but if the likelihoods at these ordered points are monotonic (or even changing smoothly), then there is structure that relates the likelihoods. Therefore, we can assume that observers can perform some degree of interpolation between smaller numbers of fixed points to arrive at a functionally much larger number of representation points.

We also assumed that observers would be able to estimate distance accurately—particularly in our experiments where the magnitudes are provided numerically—but not necessarily weight that distance in any special way. This was another motivation for using a rectangular kernel.[2] However, even the rectangular kernel needs a bandwidth, and the choice of the bandwidth may be problematic. Choosing this parameter requires a judgment about how far away a point can be to still have an effect on the subject's representation. This is another parameter that no doubt reflects individual differences and task demands. For most of our simulations, we held the bandwidth constant at 10, which seemed to work well. We could also consider an adaptive process for setting the bandwidth, perhaps based on how frequently observations fall within a bandwidth distance of representation points at different locations, but so far that has proven to be an unnecessary complication.

In the next sections we present the results of simulations designed to examine the contributions of different components of the model and then fits of the model to data from four experiments. The model produces sequences of responses that are very similar to the responses produced by human observers.

Simulations

Our simulations had three objectives. First, we wished to establish that the model approximates optimal behavior under changes in stimulus discriminability and signal frequency. That is, we wanted to show that we can apply the methods of SDT analysis to the data produced by the model and obtain estimates of d' and β, estimates that show the same sensitivity to experimental conditions as estimates computed from human data.

In particular, we calculated estimates of d' and β for 1,000 simulated subjects in several simulated experimental conditions. We used the estimate of discriminability d' given by

$$\hat{d}' = z(H) - z(FA),$$

where *H* is the hit rate, *FA* is the false-alarm rate, and *z* is the inverse of the standard normal cumulative distribution. We used the estimate of β given by

$$\hat{\beta} = \exp\{-(z(H)^2 - [1 - z(FA)]^2)/2)\}.$$

Our simulations varied the separation between the signal and noise distributions and the prior probability of a signal. When discriminability increases, estimates of d' also increase, and when the probability of a signal changes, estimates of β change.

Second, we wanted to explore the contribution of the model parameters to the model's behavior. In the first simulation, we demonstrate the effect of the maximum number of points $N_{\max}$ on the model's performance. In the second simulation, we manipulate the relative frequency of signal and noise stimuli.

Third, we explored the model's behavior under changes in payoffs. We know that human observers adjust their response strategies to approximately maximize expected earnings. We assumed that under different payoffs, our model would modify the decision rule to select a different likelihood criterion. So, for example, if the payoff for a correct "noise" decision were 4 times that for a correct "signal" decision, then the signal likelihood would need to be 4 times (more or less) that of the noise likelihood to make a "signal" decision. Finally, we looked for signature

[2] Most of the simulations presented in this article were duplicated by means of a Gaussian kernel, with no change in the results.

changes in the ROC curve that typically co-occur with changes in payoffs and stimulus base rates.

Except where noted, for each simulated subject, we presented the model with a sequence of samples drawn from one of two Gaussian distributions. The model made signal–noise decisions, and the stimulus representation was updated for 340 trials (which was the number of trials in our Experiment 1).

Manipulating the Frequency of Signal Stimuli

The first simulation illustrates the effect of changes in the probability of a signal $P(S)$ on response frequency with accurate feedback. We examined the performance of the model under three prior signal probabilities, $P(S)$ = .2, .5, and .8. We expected different levels of signal probability to result in effects similar to those observed in the classic SDT model by increasing and decreasing β (reflecting a conservative and liberal bias, respectively). However, these changes do not reflect changes in the decision rule (criterion) for the model, but instead result from changes in the stimulus representations: Because with accurate feedback the joint distribution over signal and noise stimuli is tracked (up to a constant of proportionality), the marginal probabilities, including $P(S)$, are naturally also estimated.

For this simulation, the signal and noise stimuli were drawn from Gaussian distributions with means of 60 and 40, respectively, and standard deviations equal to either 10 (in the low d' condition, $d' = 2$) or 6.67 (in the high d' condition, $d' = 3$). The bandwidth h was equal to 10, $\lambda = 0.1$, and we used the updating scheme for representation points described above. Each simulated subject began with two representation points at 40 and 60, and then a new point was added at the location of each presented stimulus. When the number of representation points reached $N_{\max} = 10$, on each subsequent trial the oldest point was replaced by a new point at the location of the stimulus presented on that trial (i.e., the value of γ was 1 for these simulations).

Figure 3 shows the results. When $P(S)$ increases, the model makes more "yes" responses, and estimates of β decrease; when $P(S)$ decreases, the model makes more "no" responses, and estimates of β increase. Also, when the standard deviation decreases (and d' increases), the estimates of d' increase accordingly. Figures 3A and 3C show the ROC clusters and estimates of β, respectively, for n = 1,000 simulated subjects with $d' = 2$; Figures

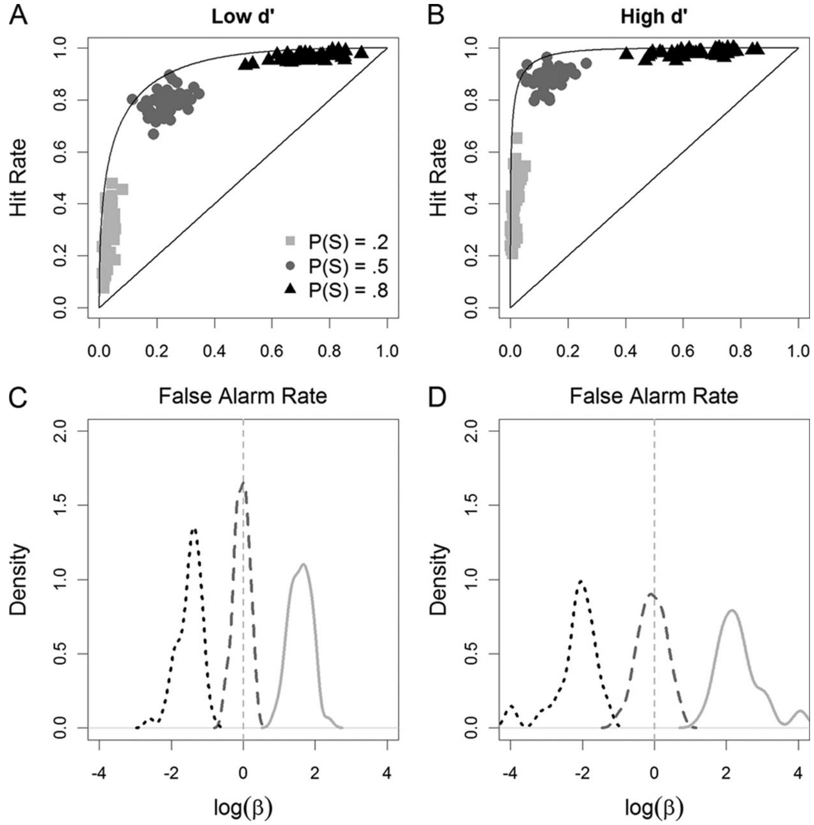

Figure 3. Changes in receiver operating characteristic (ROC) curves and estimates of β with changes in signal probability. Figures 3A and 3C show the ROC clusters and distributions of the estimates of β, respectively, for 1,000 simulated subjects performing with discriminability $d' = 2$. Figures 3B and 3D show the ROC clusters and distributions of estimates of β, respectively, for 1,000 simulated subjects performing with discriminability d' = 3. Signal frequencies $P(S)$ = 0.2, 0.5, and 0.8 are denoted by light gray squares, dark gray circles, and black triangles (A and B), respectively, and solid, dashed, and dotted lines (C and D), respectively.

3B and 3D show the ROC clusters and estimates of β, respectively, for $n = 1{,}000$ simulated subjects with $d' = 3$.

Estimated d' values (not shown in Figure 3) did not change with changes in signal frequency, but did so with changes in standard deviation. This result, together with changes in estimates of β under different signal frequencies, is consistent with published reports of human performance under similar conditions (e.g., Green & Swets, 1966). The model selects responses in the same way regardless of the value of $P(S)$, without any changes in parameter settings: It always chooses the response with the highest likelihood given the stimulus presented. If likelihood is a monotonic function of stimulus value, then this is equivalent to setting $\beta = 1$ for all values of $P(S)$. Changes in the signal frequency result in higher likelihoods for the signal representation and lower likelihoods for the noise representation (see Equations 2 and 3). This change in the relative sizes of the signal and noise representations results in estimates of β that mimic changes in a response criterion.

Manipulating Payoffs

There are two important empirical findings that our model must be able to reproduce. First, when presented with unequal payoffs for "signal" and "noise" responses, observers act to maximize their total payoffs. Second, changes in payoffs (and stimulus probabilities) influence the shape of the ROC curve in ways that the classic SDT model cannot explain. We address each of these findings in turn.

To begin, let v_H be the value or payoff for a hit and let v_{CR} be the value or payoff for a correct rejection. Assume that there are no penalties or payoffs for incorrect decisions. For equiprobable signal and noise stimuli, placing the likelihood criterion at $\beta_0 = v_{CR}/v_H$ maximizes payoffs for the equal-variance signal detection model. For our new model, the decision rule can be altered to accommodate differential payoffs by making a "signal" response only when the likelihood of the stimulus under the signal distribution exceeds that under the noise distribution by some value. With this rule, we can examine the effect of changing payoffs by looking at the mean per-trial payoffs as the critical decision ratio changes. This idea is conceptually equivalent to the response rule in classic SDT. However, the likelihoods underlying the likelihood ratio computation are the dynamic estimates constructed by our model.

We examined mean per-trial earnings for each of five v_{CR}:v_H payoff conditions: 4:1, 3:2, 2.5:2.5, 2:3, and 1:4. For each payoff condition and response rule, we simulated 10,000 responses to equiprobable signal and noise trials (with means of 40 and 60 and standard deviation 10) for 1,000 subjects following the procedure described in the previous section. The model parameters were constant across simulations with $\lambda = 0.1$ and bandwidth $h = 10$. For each simulated subject, $\gamma = 1.0$ and $N_{\max} = 30$. We used a very large number of subjects and observations and chose a large value for $N_{\max}$ to evaluate the best possible earnings of the model. Figure 4 shows the mean per-trial earnings for the model (gray) and the ideal observer SDT model (black) as a function of likelihood ratio for each payoff scheme. The standard errors were very small, and so they are omitted from the figure.

The model maximizes payoffs at approximately the same likelihood ratio as the ideal observer SDT model regardless of the value of λ. The smaller λ is, the greater the asymptotic fidelity of the representations will be, so larger values of λ, although they may result in lower overall earnings, still produce maxima at

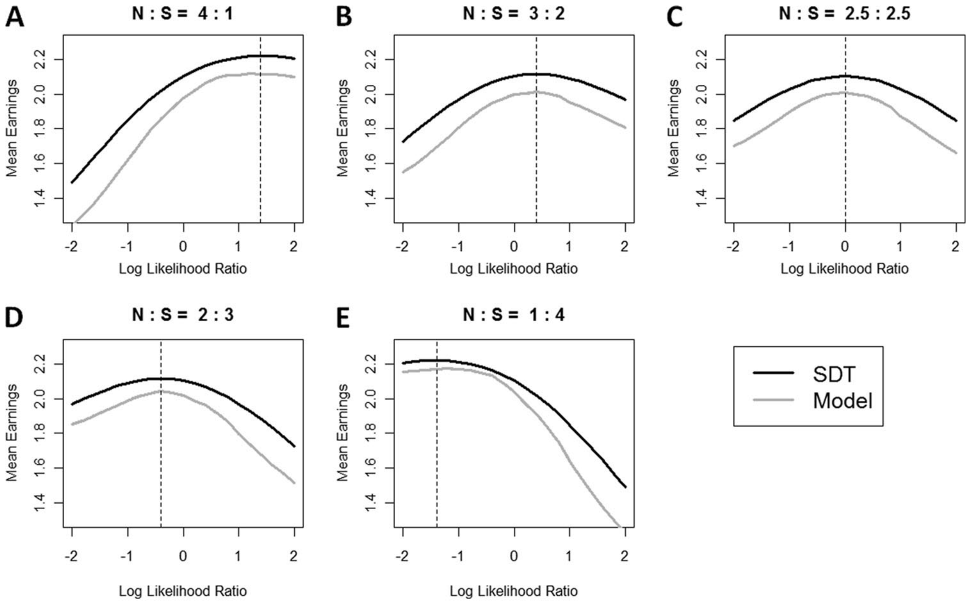

Figure 4. Mean per-trial earnings with changes in payoffs as a function of the log-likelihood ratio response rule. Each figure shows a different payoff scheme (4:1, 3:2, 2.5:2.5, 2:3, and 1:4 correct rejection N to hit S payoffs). The expected earnings for the signal detection theory (SDT) ideal observer are shown in black, and the mean earnings for the model are shown in gray. The ideal observer's criterion is shown as a vertical dashed line.

approximately the same location as the ideal observer. The effects of γ are similarly very small because of the large value for $N_{\max}$. Unlike the ideal observer, the model is unable to achieve the theoretical maximum payoff because of the imperfect nature of the stimulus representation. Comparisons between the model's performance and the theoretical maximum would then lead one to conclude that the simulated observers were not moving their criteria far enough in the direction of the bias evoked by the payoff scheme. That is, for the 4:1 payoff scheme, for example, the maximum earnings for the model are lower than those for the ideal SDT observer, indicating that too few "noise" responses are made by the model or that the SDT criterion was not placed far enough to the right. This reproduces the classic finding that human observers, also, do not seem to use criteria that are sufficiently extreme under unbalanced payoffs (Green & Swets, 1966; Healy & Kubovy, 1981).

The model's failure to attain the maximum earnings under equal payoffs (2.5:2.5) is also consistent with findings that observers violate static cutoffs even after considerable experience with the task (Erev, 1998; Kubovy & Healy, 1977). That is, although the model's payoffs were maximized by the decision rule to respond "signal" whenever the likelihood ratio was greater than 1, the mean per-trial earnings were still less than the theoretical optimum. If this result was interpreted from a classic SDT standpoint, we would have to conclude that on some proportion of the trials the observer violated the decision rule, responding "noise" when the likelihood ratio was greater than 1, or vice versa. This is one of the findings that led to speculations about criterion noise. Our model explains this finding through changes in the representation over time, which lead to reversals in the likelihood ratios and hence apparent violations of a static criterion.

Classic SDT predicts that with no changes in the underlying representation, changes in payoffs and stimulus base rates should result in movement of the hit and false-alarm rates along a constant ROC curve determined by the means and variances of the stimulus representations. This means that if the ROC curve is transformed to normal coordinates, the slopes and *y*-intercepts of the *z*ROC curves under different payoffs and base rates must be constant. However, this does not usually occur. Changes in payoffs and base rates also result in changes in the slopes and intercepts of the *z*ROC curves (e.g., Mueller & Weidemann, 2008; Van Zandt, 2000). Increases in bias to respond "signal" tend to increase the slope of the *z*ROC, although the effects on the slope are smaller when they are induced by asymmetric payoffs than when they are induced by changes in signal frequency (Healy & Kubovy, 1981).

In another simulation similar to the one just described, we factorially varied payoffs and base rates. We collected "confidence ratings" from the model on a 6-point scale, where the level chosen (1, 2, 3, etc.) depended on the perceived likelihood ratio at the representation point nearest a presented stimulus. The effect of changes in base rates was to change the likelihood ratio cutoffs, which we set to $(0.1, 0.5\beta_0 + 0.05, \beta_0, 2.5 + 0.5\beta_0, 5.0)$, and recall that $\beta_0 = v_{\mathrm{CR}}/v_{\mathrm{H}}$. (The placement of the criteria did not affect the shapes of the *z*ROCs.) For each simulated subject in each condition, we computed the slope and intercept of the *z*ROC curve, and the estimated slopes and intercepts of the *z*ROC for one parameter combination are presented in Figure 5.

Figure 5 shows the slopes and intercepts of the *z*ROC curves estimated for each subject with $\lambda = 0.1$ and $N_{\max} = 10$ over five payoff conditions and three signal base rates. The differences in the slopes and intercepts of the best fitting *z*ROC lines are considerable across base rates but relatively insensitive to payoffs. To explore the effects of λ, γ, and $N_{\max}$, we conducted a series of simulations in which all three parameters varied factorially. We found that with different combinations of λ, γ, and $N_{\max}$, we could change the size of the effect of base rate on both slopes and intercepts, as well as the ordering of the slopes and intercepts with base rate. Nothing, however, changed the size of the effects of payoffs, which remained relatively much smaller than the effects of base rate.

The most consistent empirical result, that base rates tend to have a larger effect on the *z*ROC curves than payoffs, is therefore reproduced by our model (Healy & Kubovy, 1981; Van Zandt, 2000). Healy and Kubovy (1981) showed that effects of probability are larger than effects of payoffs on criterion placement even when the bias dictated by the payoffs is larger than the bias dictated by base rates. From this effect they argued for a probability-matching mechanism that could be modulated somewhat by payoffs.

Improving the Support Representation

We explored the model's behavior in three additional conditions in which $N_{\max}$ equaled either 5, 7, or 10. For this simulation we

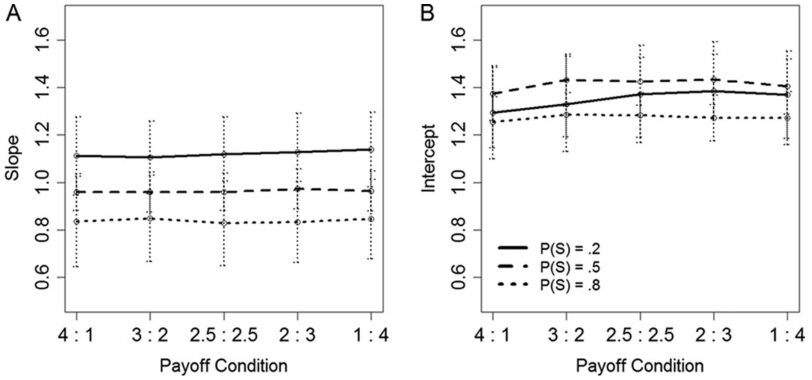

Figure 5. Effects of payoffs and stimulus base rates on *z*ROC slopes (A) and intercepts (B) generated by the model with $\lambda = 0.1$ and $N_{\max} = 10$.

held $\lambda = 0.1$, $\gamma = 1.0$, and the bandwidth $h = 10$. The probability of a signal $P(S)$ was always 0.5, and the standard deviations of both signal and noise distributions were 10. Figure 6 illustrates the effect of increasing the number of representation points. Figure 6A shows the ROC cluster for 1,000 simulated subjects for each condition. Squares, circles, and triangles indicate the performance of simulated subjects for five, seven, and 10 representation points, respectively. The solid line on the ROC curve is the performance of the ideal observer ($d' = 2$). Figure 6B shows the estimated distributions of β, and Figure 6C shows d' over all simulated subjects. As the number of points increases, estimates of discriminability increase, which is consistent with a richer, more detailed stimulus representation. The density of the representation had little effect on bias: All simulated subjects had estimated bias close to 1. (Increasing $N_{\max}$ past 10 had little additional effect on estimates of d' and β.)

The maximum number of representation points $N_{\max}$ makes an important contribution to the performance of the model. When the number of points is small, each point will be on average farther from the presented stimulus, and therefore the representation may not accurately reflect the likelihoods for that stimulus. This reduces the discriminability of signals and noise, which results in lower estimates of d'. As the density of points increases, more points will be located in the central range of the stimulus distributions and potentially closer to any selected stimulus, resulting in stimulus representations with higher fidelity.[3]

Summary

In the traditional SDT paradigm, changes in the statistical properties of the stimuli (such as discriminability and signal frequency) result in changes in classic SDT model parameters. Increased discriminability results in increases in d', and changes in signal probability result in shifts of the criterion β. Changes in payoffs result in changes in β, consistent with an attempt to maximize total earnings. In the dynamic model, we see changes in performance that are consistent with the traditional SDT interpretation, but in each simulation there were no changes between conditions in any of the model's parameters. In situations where payoffs change, the decision rule—a simple evaluation of likelihoods—can also be changed to maximize payoffs. Changes in the ROC curve under bias manipulations can also be explained by changes in the stimulus representation. All things being equal, changes in performance arise from changes in the stimulus representation, which are driven by the stimulus stream.

There are only four critical parameters of the dynamic model: the maximum number of representation points $N_{\max}$, the bandwidth h, the point updating probability γ, and the attention weight λ. The parameters $N_{\max}$ and λ contribute to the stability and richness of the stimulus representation. In static environments, stable representations (larger values of $N_{\max}$ and smaller values of λ) lead to better performance and higher d'. Changes in bandwidth did not contribute to changes in performance in the environments we used here. Later, we show that bandwidth may be important in situations that (from an SDT standpoint) require multiple criteria (see Experiment 3).

Experiments and Model Evaluation

We now present the results of four experiments designed to evaluate the model. In each experiment, subjects were presented with a number between 1 and 100 and asked to make a decision about whether it was a signal or noise. The model assumes that the subjects' stimulus representations will be built over time and that, over time, these representations should come to closely resemble the S and N distributions from which the number stimuli are sampled. Our focus is on how well the model can explain changes in the subjects' response performance over time.[4]

The first experiment is a simple signal detection task with two discriminability conditions. The second experiment is similar, except that we used different stimulus distributions and changed the prior probability of a signal over blocks of trials. The third experiment used stimulus distributions that overlapped in such a way that a traditional SDT interpretation would assume that observers maintained multiple criteria. The fourth experiment required observers to respond without feedback and with an abrupt change in the stimulus stream.

Experiment 1

Method. Experiment 1 was originally conducted by Van Zandt and Jones (2011) for a different purpose. Subjects were told that a deadly disease was infecting the people in a community. Infection was detectable by a blood assay, which returned results in the form of a number between 1 and 100. Uninfected patients had lower assay values than infected patients, and the subjects' task was to decide, for each assay, whether the patient should be treated for the disease. A mistake, either in failing to treat an infected patient or in treating an uninfected patient, resulted in the patient's death. Feedback about whether the patient lived or died was provided after every decision.

The assay values were drawn from distributions with means of 40 and 60 for well and sick patients, respectively. The standard deviation of the distributions ($\sigma = 6.67$ or 10) determined the discriminability of the two patient populations ($d' = 3$ or 2, respectively). Subjects performed 340 trials over five blocks under either high or low discriminability conditions. See Van Zandt and Jones (2011) for additional methodological details.

Results and discussion. Each subject's hit and false-alarm rates are shown as gray circles in Figure 7. Subjects in the high discriminability condition (Figure 7A) had higher hit rates and lower false-alarm rates than subjects in the low discriminability condition (Figure 7B). Shown with the subjects' data are the results of simulations conducted for each subject with the same sequence of stimuli that the subjects experienced over the 340 trials. There were no free parameters adjusted across the simulations, for either different experiment conditions or different subjects. The attention weight λ was equal to 0.08, $N_{\max} = 10$, $\gamma = 1.0$, and the bandwidth $h = 10$.

[3] The results were the same when, using fixed points, we decreased the bandwidth with increasing point density, keeping the average number of points updated on each trial equal to 1.

[4] Because ROC and *z*ROC curves are constructed by collapsing the data over time, we will not consider further the shapes of the ROCs and *z*ROCs in our experiments.

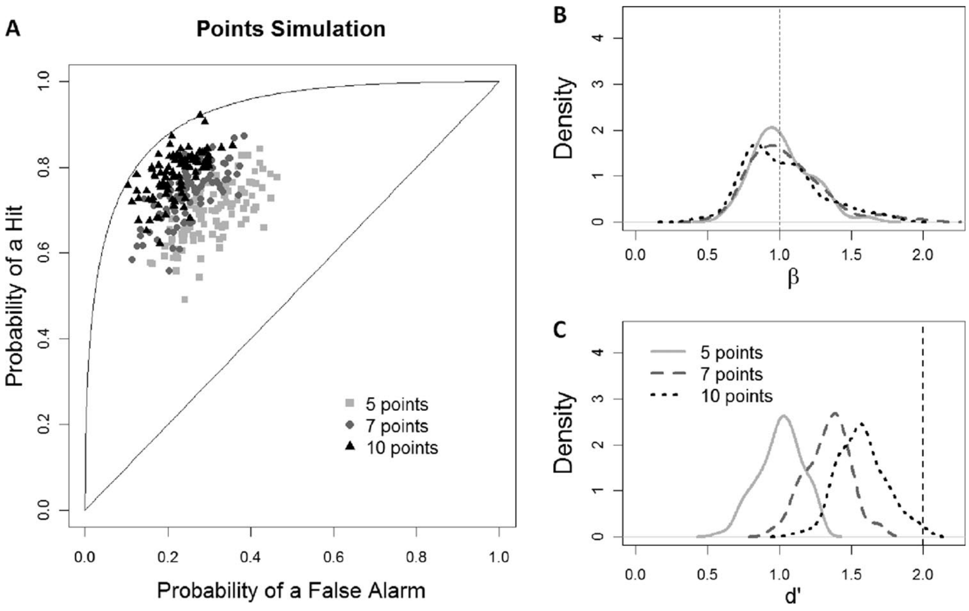

Figure 6. The effects of maximum number of representation points on response frequency. Figure 6A shows the receiver operating characteristic cluster of 1,000 simulated subjects for each group (five, seven, and 10 points are shown as squares, circles, and triangles, respectively). Figures 6B and 6C show the distributions of estimated bias β and estimated d', respectively, for each simulated subject.

The simulations began with a uniform prior over two representation points at 40 and 60, which were then updated following Equations 2 and 3 as if the effect of instructions, which stated that well patients had average assay values of 40 and that sick patients had average assay values of 60, were the same as if we had presented the stimulus 40 with feedback "noise" and then 60 with feedback "signal." The model then responded "yes" or "no" to each stimulus in a simulated sequence depending on which estimated likelihood—$\hat{f}_{S,n}(x_r; h)$ or $\hat{f}_{N,n}(x_r; h)$—was higher at the representation point x_r closest to the presented stimulus y_n. The likelihoods were then updated, also following Equations 2 and 3. Representation points were added until the number of representation points reached $N_{\max}$, at which time the oldest point was replaced with a new representation point at the value of the presented stimulus ($\gamma = 1$). For each sequence of stimuli the model generated a hit rate and false-alarm rate, and these points are shown in Figure 7.

Decreases in d' resulted in decreases in the hit rates and increases in the false-alarm rates for the simulated and the observed data, as shown in Figure 7. The model points match the observed

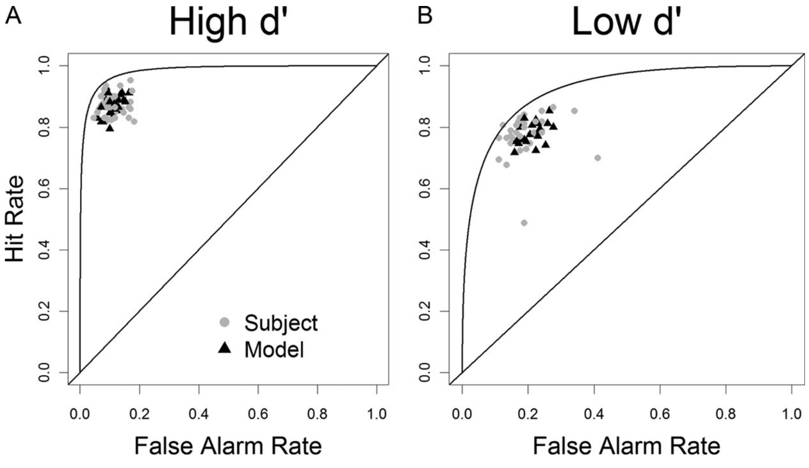

Figure 7. False-alarm and hit rates for each subject (gray circles) and simulation (black triangles) in high d' (A) and low d' (B) conditions. The simulations used the same stimulus set that the subjects experienced. The receiver operating characteristic curve in each figure (solid line) represents the optimal observer under classic signal detection theory.

points in both location and variability for both discriminability conditions.

Experiment 2

As we showed in our simulations, our dynamic model builds stimulus representations that adapt to changes in the statistical characteristics of the stimulus environment. This adaptation is gradual; stimulus representations become accurate only after some experience. This means that there should be some inertia in performance after a change to a new environment (with new statistical properties), because representations appropriate for old environments will not be appropriate for new environments.

In Experiment 2, unannounced changes in the relative frequencies of signals and noise occurred four times over five blocks of trials. We expected to see the frequency of "yes" responses move gradually toward the new signal frequency after each change. We simulated the model using the same stimulus sequences experienced by the subjects and plotted the change in the response frequencies from the model over trials. To provide the model with a greater challenge, we sampled stimuli from three distributions, manipulated between subjects: the Gaussian, exponential, and uniform distributions (the mean and variance of the signal and of the noise distributions did not change when distribution shape changed, nor did discriminability). We were primarily interested in whether the model could adapt to changes in the stimulus environment in the same way the human subjects did.

Method. The method for Experiment 2 closely follows that of Van Zandt and Jones (2011).

Subjects. Forty-nine naive subjects from The Ohio State University undergraduate subject pool served in this experiment in exchange for course credit. Two subjects were eliminated for failing to follow instructions. Subjects were fluent English speakers and reported normal or corrected-to-normal vision. They were randomly assigned to one of three distribution conditions (16 subjects in the first, 16 in the second, and 15 in the third).

Stimuli and apparatus. Stimuli were two-digit numerals presented with ASCII characters on a CRT computer monitor controlled by an Intel-style microcomputer located in a well-lit room. Characters were light on a dark background and presented in the center of the screen. Subjects were seated a comfortable distance from the display (approximately 1 m), with the index finger of each hand located on the *Z* or slash key on the computer keyboard. All responses were made by pressing one of these two keys.

A stimulus "window" was visible on the screen at all times. This window consisted of two vertical bars 3 screen rows high. The bars were constructed from the ASCII character "|" and were 9 columns apart. The two-digit stimulus appeared in the center of the middle row inside the window. Feedback consisted of either a string of four *O*s or the word *DIED*.

Procedure. Subjects were provided with instructions on the computer. The instructions were simultaneously read aloud by an experimenter. Subjects were informed about forthcoming experimental events, required responses, and feedback. In particular, they were told that well patients would have average assay values of 40 and sick patients would have average assay values of 60. Four sample trials were presented to illustrate the event sequence. In addition to the instructions, subjects were provided with a reminder card indicating the assignment of stimuli to response keys (which was counterbalanced across subjects).

Subjects were assigned to one of three distribution conditions. In the Gaussian condition, the *N* and *S* distributions had means of 40 and 60, respectively, with a common standard deviation of 6.67. In the exponential condition, the *N* and *S* distributions had shifts of 33.33 and 53.33, respectively, and a common rate parameter of 0.15. In the uniform condition, the *N* distribution extended from 16.91 to 63.09, and the *S* distribution extended from 36.91 to 83.09. Thus, the means and common standard deviation of the exponential and uniform distributions were equal to those of the Gaussian distributions. All samples were rounded to the nearest whole number.

Each trial began with the presentation of the two-digit stimulus, which remained visible for 100 ms. The subject's response triggered the feedback display, which was also visible for 100 ms. The intertrial interval (measured from the end of the feedback display to the onset of the next stimulus) was distributed exponentially with a mean of 400 ms and a shift of 200 ms. Thus, the intertrial interval was no shorter than 200 ms, and the exponential distribution ensured that subjects could not time or anticipate stimulus onsets.

Subjects completed five blocks of 100 trials each. Between each block, they were given feedback that indicated how many of their patients they had saved and how many had died. The frequency of sick patients changed from block to block. For all subjects, the frequency of sick patients (the number of samples from the *S* distribution) in the first block was 0.5. In the second block, this frequency shifted to 0.8. In the third block it shifted back to 0.5, then to 0.2 in the fourth block. Finally, the frequency returned to 0.5 in the fifth block.

Results and discussion. The goal of Experiment 2 was to show changes in response frequency over trials, changes that resulted from changes in the statistical properties of the stimulus environment. However, illustrating such changes is somewhat difficult: Each subject provides only a single "yes" or "no" response at each trial. We considered different ways to show the proportion of "yes" responses on every trial (including a moving average scheme), but decided to use a Bayesian estimation procedure for detecting change points in time series data (Smith, 1975). An explanation of this procedure and code to implement it are provided by Lee and Wagenmakers (2010, p. 47).

In brief, the procedure estimates the posterior distribution of the probability that the proportion of "yes" responses has changed on any trial (subject to some minimally limiting prior assumptions, such as that there are four change points in the series and that the change points occur somewhere within 50 trials of the true change). The heart of the model is a binomial likelihood (for the "yes" or "no" response) with a beta(1, 1) prior on the probability of a "yes" response. This results in a conjugate beta posterior for the probability of a "yes" response. We then constructed a hierarchical model for the location of the change point—the trial on which the probability of a "yes" response changes. Each subject's change points were assumed to be normally distributed with a common mean and variance. The prior distribution of the mean was assumed to be Gaussian with mean equal to the true change point and variance equal to 100. The prior distribution of the variance was assumed to be gamma(.01, .01)—an uninformative prior. We then estimated the posterior distributions of the four change point means. The estimates of this posterior give us a picture, informed by each subject's responses, of where in the series of trials changes were detected.

Figures 8A, 8C, and 8E show the proportion of "yes" responses across subjects (black) computed at every 10th trial for the Gaussian, exponential, and uniform distribution conditions, respectively. Each figure also shows the proportion of "yes" responses across simulated subjects (gray). We simulated the model using the stimulus sequences experienced by each subject, and $N_{\max} = 20$, $\lambda = 0.66$, and $\gamma = 1.0$. These parameter values produced good matches to the data; in particular, the increased value of λ relative to Experiment 1 reflected the need to adapt more quickly to the dynamic environment of Experiment 2. The regions shown around the computed proportions are the 95% Bayesian credible intervals for the beta posterior distribution of the probability of a "yes" response. Figures 8B, 8D, and 8F show the estimated posterior distributions of the mean locations of the four change points for both the real (black) and simulated (gray) subjects.

Both the real and simulated subjects varied the proportion of "yes" responses across trials to approximately match the proportion of signals in the environment, although there is some variability across distribution conditions. Subjects were able to make reasonably accurate discriminations between signals and noise, regardless of the distribution from which the stimuli were sampled. The 95% Bayesian credible intervals for the estimated d's over subjects were (2.08, 2.22), (2.29, 2.43), and (1.98, 2.12) for Gaussian, exponential, and uniform distributions, respectively, with an assumed Gaussian prior of mean 0 and standard deviation 2.

The modes of the change point posteriors are very close to the trials at which the change occurred (shown as the vertical lines at 100, 200, etc., in Figures 8B, 8D, and 8F). The range of the posteriors is also quite narrow across all conditions, rarely exceeding 10 trials in either direction. This indicates that, overall, the subjects and the model were quite fast to detect a change. Very sharp, precise posteriors indicate that most subjects perceived the change at approximately the same time. For example, this occurred around Trial 100 for the Gaussian condition and around Trial 101 for the uniform condition. Similarly, more disperse posteriors indicate that not all subjects perceived the change at the same time.

Because the perceived change point depends on the randomly generated sequence of stimuli, some subjects may vary their response frequencies in such a way that it might appear as though a change had been detected even before such a change was implemented. Therefore, we expect to see some probability mass to the left of the true change point. However, we expect that, generally, subjects will not detect a change until after it has actually occurred. The area under the posteriors from the change points (100, 200, etc.) to positive infinity is the probability that the subjects (and model) perceived a change later than the point when the change was introduced. This probability can therefore be used as a mea-

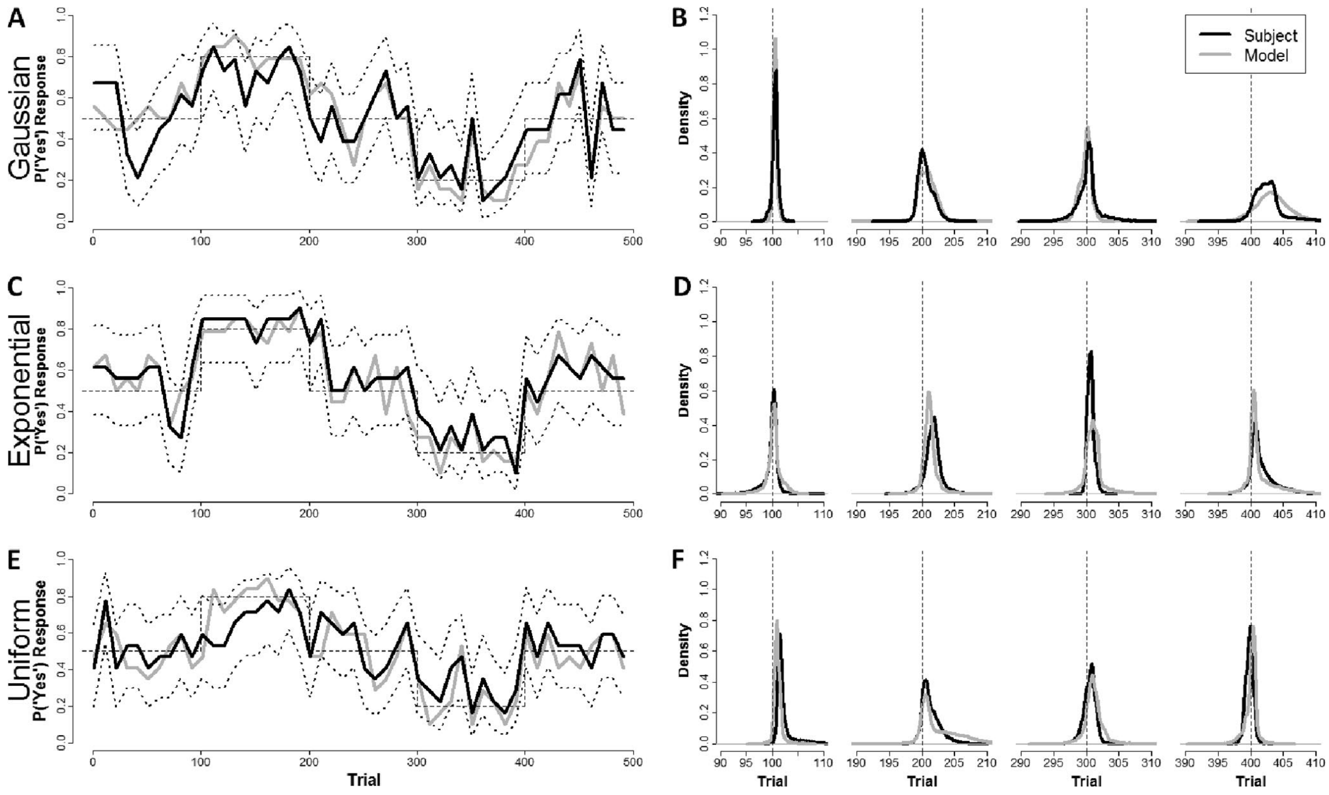

Figure 8. Probability of making a "yes" response over trials (A, C, and E) and posterior change point distributions (B, D, and F). The data from the subjects are shown in black, and the data from the simulations are shown in gray. Error bars around the data (dotted lines; A, C, and E) show the 95% Bayesian credible interval for the mean every 10 trials. The Gaussian (A and B), exponential (C and D), and uniform distribution (E and F) conditions are shown. The dashed horizontal lines (A, C, and E) show the true signal frequency.

sure of how quickly subjects (and the model) were able to adapt to changes in the stimulus sampling scheme.

Table 1 shows the expected mean change point $E[T]$ and the probability that the mean change point T is greater than the actual change point. Overall, the model follows the undulations in the observed data remarkably well, and the effects of different distributions (which are small and inconsistent) are captured in the simulations. The model seems no better at detecting the change than the subjects. In some conditions, the model's expected mean change point occurs later than the subjects', but in other conditions this is reversed. In some conditions, the model's probability of the mean being greater than the true change point is smaller than the subjects', but in other conditions this is reversed. Recall, however, that we have not adjusted any parameters to obtain the results in Figure 8. Parameter adjustments to account for individual differences would result in model behavior that even more closely approximates the data.

Experiment 3

Experiments 1 and 2 demonstrated that the dynamic model can produce data that look remarkably similar to data from subjects experiencing changes in stimulus discriminability and signal frequency. Experiment 3 examined the model's ability to accommodate behavior in a task that, under classic SDT, requires the application of more than one decision rule. Because the model does not require the specification of a fixed criterion, it should be well suited for such a stimulus environment.

In Experiment 3, the S and N distributions were Gaussian, and both had means of 50. The standard deviation of the S distribution was larger than the standard deviation of the N distribution. Thus, very low and very high stimuli were likely to be signals, whereas stimuli closer to 50 were likely to be noise. The likelihood ratio, therefore, decreased for stimuli less than 50 but increased for stimuli greater than 50. Previous research in categorization demonstrates that, though difficult, people can make such discriminations after some practice (Ashby & Maddox, 1992; Nosofsky & Stanton, 2005; Ratcliff & Rouder, 1998; Rouder & Ratcliff, 2004). The dynamic model, because it incorporates the statistical properties of the presented stimuli, will also be able to make such discriminations and, in addition, make them as the stimulus frequencies change over trials.

Method. The method, except where noted, was identical to that of Experiment 2.

Subjects. Sixty-eight naive subjects from The Ohio State University undergraduate subject pool served in this experiment in exchange for course credit. Ten subjects were eliminated for failing to follow instructions.

Procedure. The procedure was identical to that of Experiment 2, except that subjects were told that both well and sick patients had average assay values of 50 but that sick patients were more likely to have extreme assay values. The N and S distributions from which stimuli were sampled had a common mean of 50. The standard deviations for the N and S distributions were 2.13 and 15.00, respectively. As in Experiment 2, the frequency of sick patients changed from block to block.

Results and discussion. We simulated data from the model by presenting it with the same stimulus sequences experienced by the subjects. As in Experiment 2, the attention parameter was fixed at $\lambda = 0.66$, $N_{\max} = 20$, and $\gamma = 1.0$. However, we decreased the bandwidth to $h = 3$. This change was necessary because of the narrowness of the N distribution. The smaller bandwidth ensures that for stimuli from the N distribution, the N representation is updated for points in the same critical region.

Figure 9A shows the frequency of "yes" responses as a function of stimulus magnitude averaged across subject–simulation and trial. The two vertical dashed lines represent the points at which the true likelihood ratio changes from being greater than 1 (at 45.75) to less than 1 and then again to greater than 1 (at 54.25); these are the two points at which an SDT optimal observer would place decision criteria. The dashed line shows the probability that a stimulus is a signal given its magnitude. The probability of a "yes" response for the subjects (black line) is very similar to the probability of a "yes" response for the simulations (gray line). In particular, neither the subjects nor the model simulations matched the probability that the stimulus was a signal: The probability of a "yes" response—for both observers and the model—was much

Table 1
Estimates of the Expected Value of the Change Point T and the Estimated Probability That the Change Point T Is Greater Than the True Change Point (Change) for Each Condition in Experiments 2 and 3

	Experiment 2							
	Gaussian		Exponential		Uniform		Experiment 3	
Change	Data	Model	Data	Model	Data	Model	Data	Model
				$E[T]$				
100	100.63	100.56	100.33	99.80	101.08	102.13	102.95	103.12
200	200.64	200.89	201.24	201.72	202.68	201.37	203.26	202.25
300	300.30	299.89	301.10	300.71	301.08	300.50	304.77	304.42
400	402.66	403.18	401.37	401.60	400.09	399.84	403.52	404.03
				$P(T > \text{Change})$				
100	0.89	0.94	0.69	0.61	0.97	1.00	0.98	0.99
200	0.70	0.74	0.93	0.93	0.93	0.91	1.00	0.99
300	0.64	0.54	0.87	0.95	0.87	0.76	0.99	0.99
400	0.94	0.90	0.87	0.89	0.69	0.44	1.00	0.99

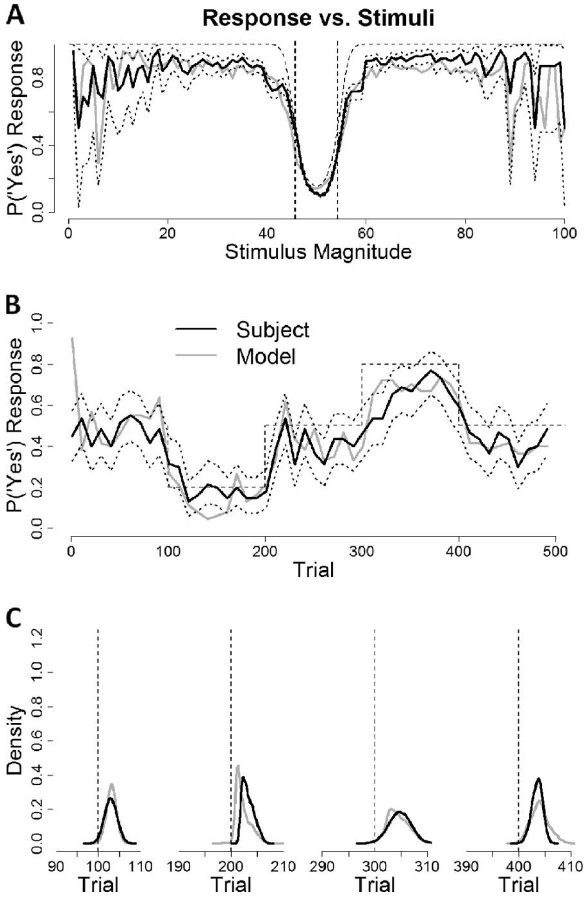

Figure 9. Figure 9A shows the frequency of "yes" responses as a function of stimulus magnitude averaged across real (black lines) and simulated (gray lines) subjects, with dotted 95% credible intervals. The vertical lines at magnitudes 45.75 and 54.25 represent the points at which the slope of the likelihood ratio changes sign. Figure 9B shows the proportion of "yes" responses every 10 trials together with the 95% Bayesian credible interval. The dashed horizontal lines show the true signal frequency. Figure 9C shows the estimated posterior mean change point distributions. The posteriors for the subjects are shown in black, and the posteriors for the simulations are shown in gray.

lower than the actual probability of a signal for extreme stimulus magnitudes (outside the region around 50).

Figure 9B shows the proportion of signal responses over trials, and Figure 9C shows the posterior mean change point distributions with the procedures described for Experiment 2. In contrast to the response profile seen for Experiment 2, the subjects in Experiment 3, on average, seem less able to adjust their responses to match the frequency of signals in the environment. The model suffers from the same deficiency, as shown by the response profile for the simulated subjects in Figure 9B. Similarly, the mean change point posteriors are less precise than those of Experiment 2, and all the probability mass lies to the right of the true change point (see Table 1). These results, together with the expected mean change points given in Table 1, suggest that neither the subjects nor the model adapts as quickly to change as in Experiment 2.

Another difference between these results and those for Experiment 2 is the lack of fit between the subjects' and the model's mean change point posteriors. However, as for Experiment 2, we did not vary the model parameters for the individual subjects. Fitting the model to account for individual differences would improve the match between the subjects' and the model's estimates of the mean change point posteriors.

For Experiment 3, the classic SDT model, absent any additional theoretical structure to explain how criteria might move from block to block, predicts that subjects should move their criteria closer together when signals are more frequent and farther apart when signals are less frequent. Our results are consistent with this explanation. However, note from Figure 9 that in all conditions, the probability of a "signal" response, from either the model or the subjects, is well below its ideal value. This occurs because stimulus magnitudes around 40 and 60 were often classified as noise, even though the probability that such stimuli arise from the noise distribution is very small. This means that if we try to explain the results using SDT, we must assume that subjects consistently place their criteria too far apart—a persistent and unexplained bias to say "noise," which does not arise from any characteristic of the stimulus representation or the reward structure of the task.

Recognition Memory and the Role of Feedback

As we outlined above, most categorization and psychophysical experiments provide feedback for at least a training period, if not the entire experimental session. By contrast, memory experiments usually do not provide feedback. Therefore, to apply our model to memory data, we must consider alternatives to the updating scheme in our model, which is based on the availability of accurate feedback after every response.

We took the easiest possible approach and assumed that subjects use their own response as the basis for feedback. That is, in the absence of better information, subjects assume that their responses are correct. This changes the updating rule from updating the representation identified by feedback to updating the representation identified by the response. As we discussed earlier, some problems arise with this scheme: most importantly, the possibility that after many responses of one kind, the representation for that response dominates the alternative representation, which then decays away to zero. Judicious selection of parameters, particularly of λ and γ, together with at least minimally informative priors, prevents one representation from dominating the other. Therefore, although we explored other options, we decided to stay with the simpler response updating rule.

In this section, we present the results from an experiment without feedback very similar to Experiment 2 and an application to the recognition memory data of Brown, Steyvers, and Hemmer (2007). We show that the dynamic model with updating of the representation associated with the response can explain the results from these two experiments. We also discuss the application of the model to recent data from Mickes, Hwe, Wais, and Wixted (2011) and Cox and Dobbins (2011), which show that it is difficult to explain some recognition memory data with the idea that subjects set fixed criteria along an axis of memory strength.

Experiment 4

A critical issue for situations where there is no feedback is how subjects respond to unexpected shifts in the stimulus environment.

With feedback, we saw in Experiments 2 and 3 that people can adjust very quickly. Without feedback, however, we assume that people adjust their response strategies much more gradually. After having built strong representations at one location on the stimulus support, subjects without feedback should be very reluctant to adjust those representations until they have seen many stimuli that are inconsistent with them. It is also possible that subjects never adjust their representations if they do not get feedback, especially for unfamiliar tasks such as the assay task we have used to this point.

Subjects in this experiment were randomly assigned to feedback and no-feedback conditions. To determine how quickly subjects adjust to changes in the stimulus stream, we shifted the means of the distributions from which stimuli were sampled halfway through the experimental session. The initial instructions to the subjects indicated the overall means of noise and signal stimuli for the entire experimental session, but told them nothing about how the means would change over the first and second halves of the session.

Method. The method, except where noted, was identical to that of Experiment 2.

Subjects. Forty-nine subjects served in this experiment. Twenty-three were randomly assigned to the feedback condition, and the remaining 26 were assigned to the no-feedback condition.

Stimuli and procedure. Subjects in the feedback condition were given the same instructions as subjects in Experiment 2, except they were not told how many patients they had saved after every block. Subjects in the no-feedback condition were not told, neither after trials nor in rest periods, whether their patients had died nor how many. Only at the end of the experiment were they told how many patients had died and how many they had saved.

All assay values were drawn from Gaussian distributions with standard deviations equal to 6.67. Sometimes the means of both the sick and well patient distributions were below 50, and sometimes they were both above 50. The means were either 25 and 45 or 55 and 75 for well and sick populations, respectively. The means of the sick and well patient populations shifted by 30 in the middle of the experiment (Trial 251) to the high means (if the initial means were low) or to the low means (if the initial means were high). Over the 500 trials, the means of the well and sick patients were 40 and 60, as indicated in the instructions to the subjects. The proportion of sick patients was always 0.5. Twenty-five subjects were randomly assigned to the increasing mean condition, and the remaining 24 subjects were assigned to the decreasing mean condition.

Results and discussion. Before data analysis, we reversed the stimuli and responses for the decreasing means condition, which then allowed us to collapse across the increasing and decreasing means conditions. We then considered estimates of an "effective" criterion over trials. That is, we computed the proportion of hits and false alarms over subjects on each trial and used these hit and false-alarm rates to estimate a criterion in the usual way.[5] Figures 10A and 10B show the effective criteria over trials for the feedback and no-feedback conditions. The data are shown as black lines over the simulated data provided by the model as gray lines.[6] In both conditions, the model performance mimicked that of the human observers.

Figure 11 shows the same effective criteria as in Figure 10, but with the feedback and no-feedback conditions plotted in the same panel to emphasize the performance differences between the two conditions. Figure 11A shows the data from the human observers, and Figure 11B shows the simulated data from the model. Subjects in both the feedback (gray lines) and the no-feedback (black lines) conditions were able to respond appropriately to signals and noise, although they did not reach the optimal level of the ideal observer, which is shown in the figure as a dashed step function. The subjects in the feedback condition had effective criteria that were closer to this ideal than subjects in the no-feedback condition. After the change at Trial 251, subjects in the no-feedback condition shifted their criteria rapidly because the stimuli presented then were extremely different from any that had been presented before. The subjects in the feedback condition were better able to adjust their criteria, whereas subjects in the no-feedback condition were consistently less than optimal. The model experienced the same difficulties as the human observers.

We can conclude, not surprisingly, that subjects who received feedback updated their representations to a greater degree than the subjects who did not receive feedback. This finding is reflected in the estimated model parameters for the two conditions: The parameters that resulted in good fits to the feedback conditions were similar to the parameter values we used in the earlier simulations ($h = 10$, $\lambda = 0.08$, $N_{\max} = 6$, and $\gamma = 1.0$), whereas the parameters for the no-feedback conditions were quite different ($h = 10$, $\lambda = 0$, $N_{\max} = 35$, and $\gamma = 0.21$).

The estimated value of $\lambda = 0$ in the no-feedback condition means that the subjects appeared to be performing the task without ever updating their likelihood estimates. For these subjects, the only changes in their stimulus representations were those induced by dropping old representation points and storing new ones. In this mode, the model operates as an exemplar model with unreliable storage (e.g., McKinley & Nosofsky, 1995; Nosofsky, 1986).

The ability to store new representation points ($\gamma > 0$) allows the model to respond to the shift in the means both at Trial 251 and on the very first trial, when the presented stimuli appear much different from what the subject was led to expect by the experimenter's instructions. However, storing new representation points also brings a danger. If the maximum number of representation points $N_{\max}$ is small, it becomes quite likely that random sequences of repeated stimuli will lead to all representation points being stored as, say, "signal" (i.e., the estimated likelihood of the signal distribution is greater than that of the noise distribution at all representation points). In the absence of corrective feedback, this situation is catastrophic—the model will thereafter always respond "signal" to every stimulus presented. The larger number of representation

[5] We computed the midpoint of $z_{FA}\sigma + \mu_N$ and $z_H\sigma + \mu_S$, where z_{FA} and z_H are the z scores for the false-alarm and hit rates, respectively, computed from the inverse normal distribution function. The standard deviation σ and the means μ_N and μ_S are the standard deviation and means of the presented stimuli.

[6] The breaks in the data are caused by trials where all subjects gave the same response, resulting in an inability to compute an effective criterion without introducing a correction factor.

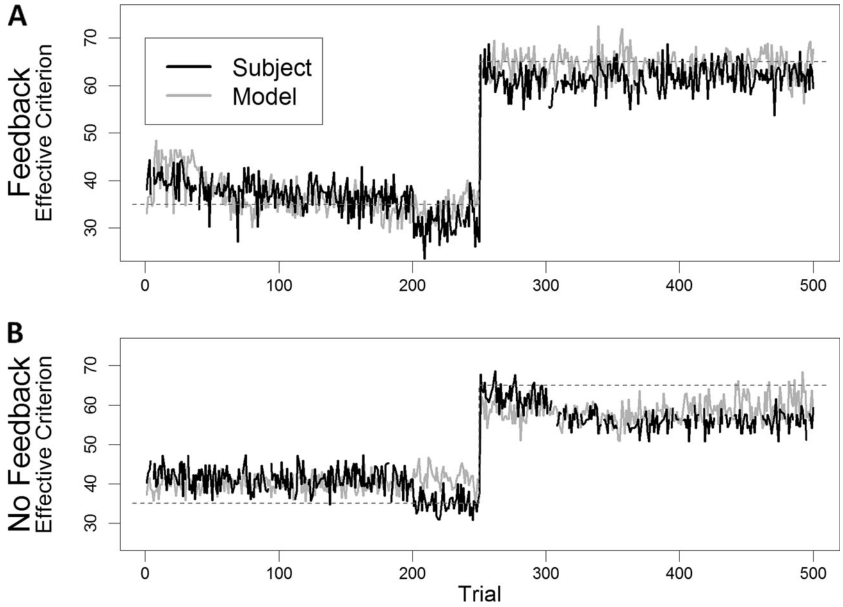

Figure 10. Effective criteria as a function of trial computed from real (black lines) and simulated (gray lines) subjects, for subjects in feedback (A) and no-feedback (B) conditions. The dotted gray step functions show the ideal criteria for a signal detection theory observer for fixed distributions at the locations described in the text.

points used in the no-feedback condition ($N_{\max} = 35$) makes this situation very unlikely.[7]

The explanations we have presented for the results of Experiment 4 depend on the existence of priors established by the experimenter's instructions. If such priors are not informative, if the experimenter's instructions do not indicate the approximate location of the signal and noise distributions, then the effects we observed under the no-feedback conditions may have looked very different from those shown in Figure 10. Therefore, we reran this experiment and eliminated the information about distribution locations from the instructions so that neither the feedback nor the no-feedback group knew the average magnitudes of noise and signals. The results were identical, indicating that subjects can establish useful representations even with noninformative priors.

Recognition Memory: The Mirror Effect

Experiment 4 required subjects to make classification decisions in a novel task with no training and no feedback. This is not typically how discrimination experiments are conducted. Feedback is almost always provided in signal detection tasks, and categorization tasks that do not use trial-by-trial feedback almost always include a training session with feedback that familiarizes the subjects with the stimuli to be presented or sporadic feedback on a subset of stimuli throughout the session. The importance of feedback in helping subjects develop a stable representation of the stimulus sets is evident in the results from our Experiment 4, where subjects who were not provided with feedback seemed not to update the likelihoods at all, and stimulus representations were "updated" by movement of the representation points only.

By contrast, recognition memory tasks almost never provide feedback, nor do they provide training. Knowledge of how to discriminate old from new items is something that subjects bring with them into the laboratory. These experiments, therefore, are examples of applications for which the SDT assumptions of established representations and fixed criteria are not unreasonable. In our model, we would instantiate prior experience in the form of well-developed, informative priors defined on some strength axis, but it turns out that the model can perform recognition memory tasks even with only minimally informative priors.

To test our model's ability to accommodate recognition memory data, we used data reported by Brown et al. (2007). Brown et al. asked participants to remember a list of pictures of objects from a well-defined category (e.g., binoculars). The subjects were subsequently tested on their ability to correctly discriminate "old" test pictures (i.e., pictures that were studied earlier) from "new" test pictures. The base rate of old pictures was 50%, but the first half of the test list was different from the second half. In the first half, the new pictures were relatively easy to identify: They were new pictures from the studied category. However, in the second half, the new pictures were much harder to identify because they were mirror images of the old pictures. Subjects were explicitly warned that mirror images of old pictures should be classified as new. Each subject performed in six sessions: three where the test lists switched from easy to difficult and three where the test lists switched from difficult to easy.

[7] The much larger number of representation points in the no-feedback condition may violate reasonable estimates of working memory capacity, but these points are not updated as in the feedback case. This means that not only are no likelihoods stored for these points as in the feedback case, but also they act as (moving) exemplars. Exemplar-based models such as Nosofsky's (1986) GCM do not generally limit the number of exemplars out of concern for memory limitations. We discuss this feature of the model in the General Discussion.

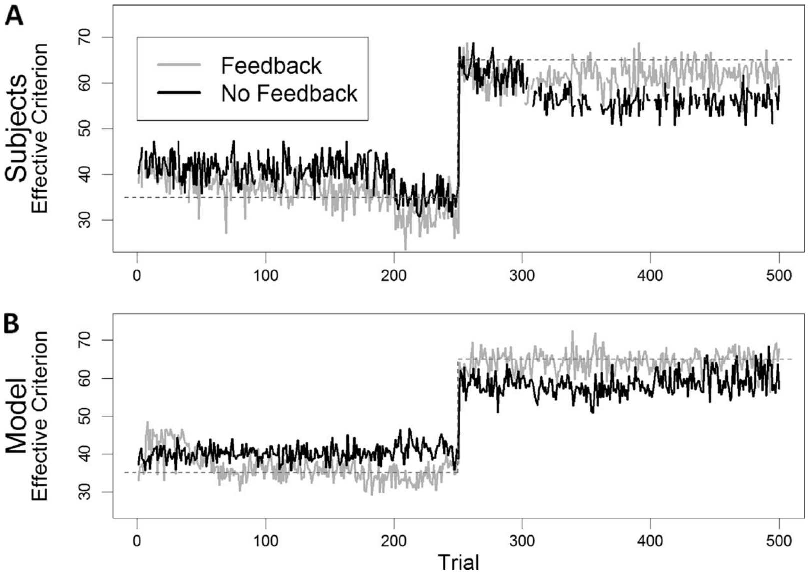

Figure 11. Effective criteria as a function of trial from Figure 10 replotted to show the contrast between feedback (gray lines) and no-feedback (black lines) conditions. Figure 11A shows the data from the real subjects, and Figure 11B shows the data from the simulated subjects. The dotted gray step functions show the ideal criteria for a signal detection theory observer for fixed distributions at the locations described in the text.

Brown et al. (2007) estimated d' for the hard and easy distractors and observed a standard mirror effect before the shift: Not only was discriminability lower for the hard distractors than for the easy distractors, but the false-alarm rates were lower and the hit rates were higher for the easy distractors (see Figure 12). However, after the shift, the mirror effect took time to reestablish. For about six trials after the shift, the false-alarm rate for difficult items increased, but the hit rate for easy items did not increase, destroying the mirror effect. Because the studied items themselves did not change during testing, changes in the mirror effect must be attributed to cognitive changes induced by context.

These data provide an interesting application for the dynamic model. First, they show a mirror effect, which our model must also be able to explain. Second, they show changes in the size of that effect over time and with changes in the stimulus stream, changes that our dynamic model must account for. Brown et al. (2007) argued for a dynamic signal detection model in which subjects changed their criterion in response to changes in the stimulus context. Because our model does not have a decision criterion in the usual sense, we must explain changes in the mirror effect by gradual changes in the estimated distributions of memory strength for old and new items.

To predict recognition memory data, our model assumes, as do most others, that each test item gives rise to an internal rating of memory strength. This rating varies randomly from item to item, but it is on average higher for old than new items, and also higher for the difficult new items than for the easy new items. As in our experiments, our model's task is to estimate the distributions of the memory strength signal for old and new items and to make responses according to the estimated likelihoods. Memory strength ratings were generated by sampling from normal distributions with unit variance. The mean strength for new items was fixed at 0. The mean strength for old items was 1.41, and for difficult new items it was 0.82.

We took the same modeling approach as for Experiment 4 and assumed that the model's response would take the place of feedback. Using the mechanisms described above and the randomly generated strength ratings, we simulated the model for 1,000 study–test blocks, all with identical parameters, 500 for the easy-followed-by-difficult condition and 500 for the difficult-followed-by-easy condition. For each simulated block the model constructed representations of old and new items, representations that gradually shifted as the new items changed. In these simulations we set $\gamma = 0.5$ and reduced λ to 0.04. Additionally, we set $N_{\max} = 12$ and the bandwidth $h = 0.40$. The prior representation was established by placing two points, one at 0.00 (the mean of the "new" distribution) and one at 0.80. We then set the prior likelihoods of the new representation as 0.75 and 0.25 for the points 0.00 and 0.80, respectively. Similarly, we set the prior likelihoods of the old representation at 0.25 and 0.75 for the points 0.00 and 0.80, respectively. This prior was reset for each new study–test cycle, in line with the experimental method that used unrelated test lists (e.g., pictures of binoculars for the first list, then fruit for the next list, and so on).

Figure 12 shows results of the simulation (gray lines) together with the data from Brown et al. (2007; black lines). Circles represent the hit rates and triangles the false-alarm rates. The hollow points (dashed lines) show the rates for the difficult-followed-by-easy condition, and the solid points (solid lines) show the easy-followed-by-difficult condition. The model's behavior is very similar to the observed data. The most important finding, the crossover in the hit rates from the two conditions in the last half of the block, is reproduced by the model, although it is of slightly smaller magnitude.

The dynamic SDT model can reproduce the basic patterns of data found in a typical recognition memory task, including changes in the mirror effect with changes in stimulus difficulty. By making the naive assumption that the observer replaces the feedback step

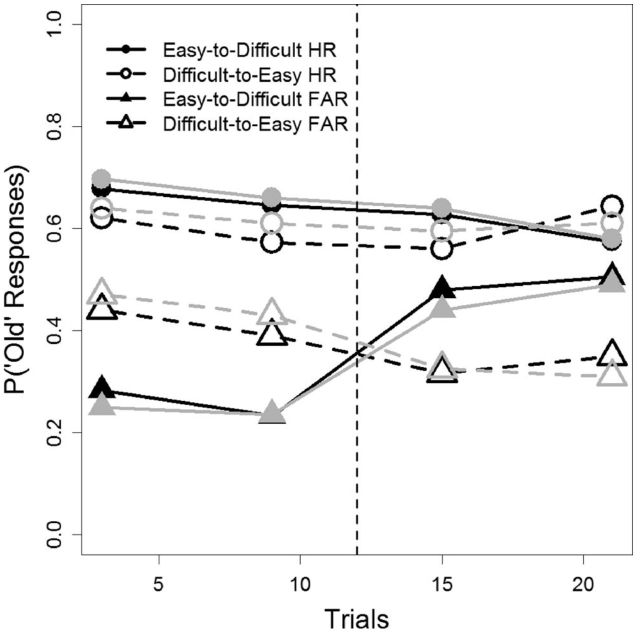

Figure 12. Frequency of "old" responses from Brown et al. (2007; black) along with the predictions of the model (gray). The triangles show the false-alarm rates (FAR), and the circles show the hit rates (HR). The hollow circles and triangles (dashed lines) show the difficult-followed-by-easy condition, and the solid circles and triangles (solid lines) show the easy-followed-by-difficult condition.

with the response (i.e., the subject assumes the response was correct in the absence of any information to the contrary), we can find parameter values that produce response rates closely matching the hit and false-alarm rates observed in Brown et al. (2007). Most importantly, no aspect of the model was changed to capture the difference between the easy and difficult experimental conditions, other than simply instantiating this difference in the simulated memory strength ratings.

Recognition Memory: Other Issues

There are two additional empirical findings in recognition memory that have been difficult to explain via classic SDT models with fixed criteria. One is the finding that with respect to confidence ratings, items with very high memory strength appear to be treated differently from all other items. The second is that recognition memory performance does not change regardless of whether distractors (or targets) are present in the test list and whether the subjects know that distractors (or targets) will be present in the test list. Our model can be adapted to accommodate both of these findings, in ways that are similar to other strength-based models. We address each finding in turn.

Strong memories are hard to scale. Mickes et al. (2011) explored the relationship between recognition memory strength and a subject's rating of that strength. Their subjects studied 150 words and then completed a recognition memory task in which they rated memory strength on a 20-point scale. Ratings from 1 to 10 indicated that they were very sure to unsure that the probe word was new, whereas ratings from 11 to 20 indicated that they were unsure to very sure that the probe word was old. The subjects were strongly discouraged from using the most extreme ratings (1 and 20), which indicated the highest possible confidence that the word was new or old, respectively.

As we described in the Confidence section of our introduction, classic SDT explains how strength is mapped into the affective judgment by placing criteria along the strength axis. If memory strength follows a Gaussian distribution, the frequency of perceived strengths must decrease as strength increases, producing the decreasing tails of the distribution. Similarly, if memory strength maps to confidence level, then the frequency of extreme confidence levels must also decrease as strength increases, producing fewer extreme ratings than intermediate ratings, unless the criterion for the most extreme ratings is set close to the means of the signal and/or noise distributions. Mickes et al. (2011) provided very strong admonitions against using the highest confidence ratings, which should have moved the criteria for these responses far into the tails of the distributions. However, they showed that in spite of these admonitions, the proportion of highest "certain old" ratings (20) exceeded the proportions of all other ratings under a variety of conditions.

Our dynamic model explains these results by way of the limited number of representation points at which likelihoods are stored. The process by which representation points are added and dropped ensures that most of the points will be located in the high-density regions of the "true" stimulus distributions. Because the number of points are limited, there will be few, if any, points in the tails of the representations. Instead, all perceived strengths greater (or less) than the most extreme representation point will inherit the likelihood information from that most extreme point. Therefore, there is no basis upon which our model would scale confidence ratings for these strong memories.

As in the simulation we conducted exploring the effects of payoffs, and following classic SDT ideas about multiple criteria, we assumed that our model makes rating judgments based on the value of the likelihood ratio at a particular representation point. Very high and very low likelihood ratios will map to the most extreme confidence ratings 1 and 20, as in Mickes et al. (2011). We simulated our model with 19 criteria, equally spaced on the log-likelihood scale, to simulate Mickes et al.'s 20 confidence ratings. The highest criterion (corresponding to a confidence rating of 20) was set at 3.7, and the lowest (corresponding to a confidence rating of 1) was set at -5. These asymmetrically placed criteria captured the observed difference between very strong and very weak memories. We then simulated 10,000 blocks of 24 trials each using the parameter settings estimated for the recognition memory data of Brown et al. (2007; see Table 2). As in the earlier simulation, the familiarity of old and new items was simulated with samples from Gaussian distributions with unit variance and means of 0 and 1.41, respectively.

Figure 13A shows the distribution of confidence ratings, and Figure 13B shows the decision accuracy as a function of confidence level. Our simulated results are very similar to the data reported by Mickes et al. (2011), particularly their Figure 5. In both our simulations and their data, accuracy is highest for the highest confidence responses, and the frequency of highest confidence responses is higher than for less extreme confidence levels. This suggests that one explanation for the failure to scale strong memories might be the sparse representation of the decision axis.

Table 2
The Parameter Values Used for Each Model Fit

Experiment	$N_{\max}$	h	λ	γ
1	10	10	0.08	1.0
2	20	10	0.66	1.0
3	20	3	0.66	1.0
4 (feedback condition)	6	10	0.08	1.0
4 (no-feedback condition)	35	10	0.00	0.21
Brown et al. (2007)	12	0.40	0.04	0.5

A prediction that our model makes, based on this same mechanism, is that weak memories should also fail to scale. However, the extent of this effect will be determined by the overlap between the target and distractor distributions.

To see how overlap will determine the extent of the failure to scale, consider the higher variance of the old-item distribution—commonly assumed in recognition memory. This higher variance ensures that there will be many more points sampled in the representation region for new items. For example, consider a new-item distribution with mean 0 and standard deviation of 1, and an old-item distribution with mean 1.5 and standard deviation 1.25. If we take the "effective range" of each distribution to be ±3 standard deviations, then samples from the new distribution will extend from -3 to 3, and samples from the old distribution will extend from -2.25 to 5.25. Over 70% of the samples from both distributions will be less than 1.5, resulting in sets of representation points that cluster around the new mean of 0. Therefore, very strong memories can span a much wider range with greater distances from the highest representation point, whereas very weak memories will not be as far from the lowest representation point. The results presented by Mickes et al. (2011) show some evidence that weak memories also fail to scale, but that this effect is smaller than for strong memories. Further experiments will be necessary to fully test this prediction.

Distractor-free performance. An acknowledged difficulty for our model is accounting for data from paradigms in which accurate feedback is not provided. The difficulty occurs during the updating process. If parameters are not selected carefully, one representation may dominate the other, leading to all "yes" or "no" responses, regardless of the stimulus presented. These difficulties may be exacerbated by high frequencies of signals or noise.

One recognition memory paradigm, explored most recently by Cox and Dobbins (2011), takes the relative frequencies of signal or noise stimulus presentations to an extreme. They asked subjects to rate how old probe words felt on a 6-point memory strength scale when test lists consisted of either all old words, all new words, or a standard 50–50 combination of both. Even when their subjects were told the relative frequencies of old words within each list type, they nonetheless produced misses and false alarms with frequencies similar to those observed in standard test conditions.

Our model might accommodate these results by again emphasizing the influence of the prior. To fit the data from Brown et al. (2007), we reduced the updating parameter λ so that representations were updated very little, if at all. As long as the priors differentiate between old and new words, even by a small amount, there will be some possibility, even with very high old-word frequencies, of making a "new" decision. That is, there will be some representation points for which the likelihood favors "new." The frequency of "new" responses will depend mostly on the prior.

The Cox and Dobbins (2011) experiment, though demonstrating that studied items elicit a range of perceived memory strengths, does not directly speak to our model, because our model explains the decision process and not the memory process. It is interesting to note, however, that in distractor-free recognition tasks, subjects are either not told that all test items are old words (e.g., Underwood, 1972; Wallace, 1978, 1982) or asked to do something different from an old–new discrimination, such as only endorsing items that they remember having seen previously (which changes the task to something like a remember–know procedure; e.g., Wallace, Sawyer, & Robertson, 1978) or estimating their confidence that they had seen the test items. For example, Cox and Dobbins explicitly instructed their subjects that there was no

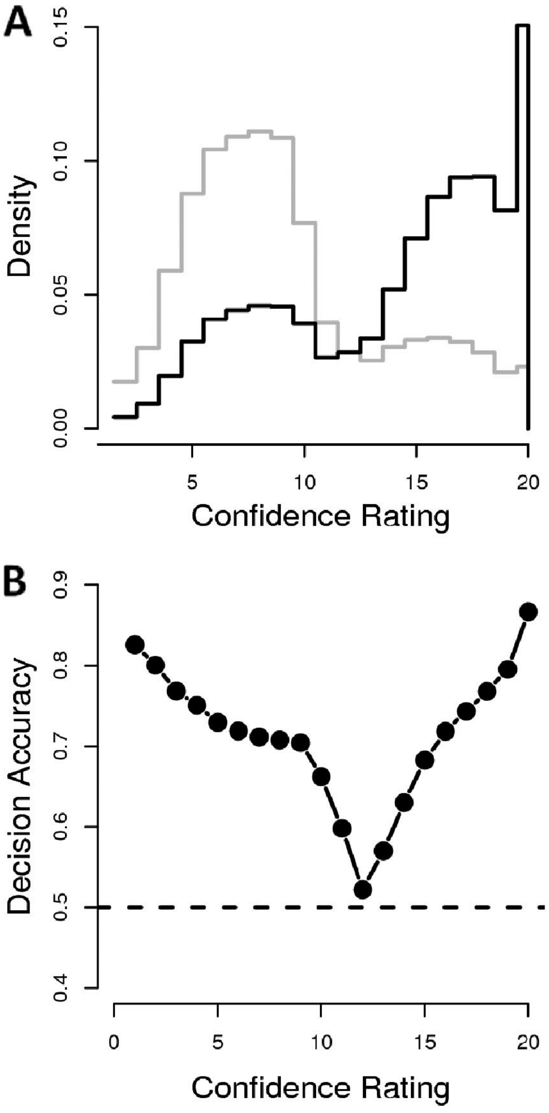

Figure 13. Frequency distributions of confidence ratings for targets (black) and distractors (gray; A), and decision accuracy as a function of confidence rating (B).

correct response to any probe item, and that instead the subjects were instructed to evaluate their feeling of how old an item was.

Having been told that all items in a test list were of one type, had the subjects been instructed to decide whether each word was old or new, and not evaluate confidence, memory strength, or whether the word was remembered, they would not produce hit and false-alarm rates that are similar to the hit and false-alarm rates in standard recognition. It is an empirical question, of course, but, arguably, not one interesting enough to warrant running the experiment: Subjects would produce much higher hit rates if they are asked to respond simply "old" or "new" to a list that they are told contains only old words (e.g., Strong & Strong, 1916), to the extent that they comply with the experimenter's instructions. The hit rate would increase even higher if incentives were provided to encourage subjects to maximize accuracy. By incorporating the information about the structure of the test list into the prior, our model would produce exactly this sort of result.

General Discussion

The classic SDT framework does not provide an explanation for how stimulus representations arise, and also assumes that such representations are relatively immutable from the first few trials of a detection task. Of course, this assumption might sometimes be well justified. For example, data might be collected only after an extensive period of practice with an independent and identically distributed sequence of stimuli, without any change in the experimental conditions, or the task could be one that relies on representations established by a subject's everyday experiences in the world. In many laboratory discrimination tasks, however, this assumption is not justified. Furthermore, the classic SDT framework assumes that subjects can accurately place a response criterion along the decision axis, and computations of statistics like d' and β assume that this criterion is constant over all the trials for which the computation is made. Although many implementations of SDT assume that the criterion changes from place to place between experimental conditions, few mechanistic accounts of this change have been incorporated into models that depend on SDT-inspired decision rules.

In this article we have proposed a dynamic SDT model that provides a detailed account of how stimulus representations develop. Our account describes how stimulus representations for signals and noise might evolve over time, and so how these representations respond to changes in the statistical properties of the environment. As a consequence of the model's changing representations, the decision rule is not dependent on a criterion location. Instead, the decision is based on the estimated likelihood ratio at a point close to the presented stimulus's magnitude. The response is given based on the likelihood that is highest: "yes" if the signal likelihood is greater than the noise likelihood (or exceeds it by some amount determined by the payoff structure) and "no" otherwise. The decision rule never has to change, but which response is most appropriate for a particular stimulus will change as the stimulus representations change. This makes strong predictions about sequential dependencies and changes in detection performance with changes in the stimulus-generating mechanism. Another benefit of eliminating the fixed criterion location is a greater ability to adapt to stimulus environments that do not have monotonic likelihood ratios.

Our model cannot replace SDT in standard applications. Our model is not appropriate for separately quantifying sensitivity and bias, a task for which classic SDT is unmatched. Our focus is on the role that SDT plays in process-level accounts of the stimulus representation that might be incorporated into larger cognitive theories. Our contribution is then to explain how humans might appear to set a decision criterion in a relatively optimal manner and change that criterion over time, even though our explanation includes no such elements (see also Glanzer et al., 1993, for a related approach confined to the recognition memory paradigm).

Under certain conditions, our model will converge[8] to the classic SDT model. In particular, suppose an observer experiences a long string of independent and identically distributed stimulus magnitudes, and $\lambda = 1/n$, where n is the number of trials. For our applications, we used a constant λ, so the representations were permitted to change continuously throughout a sequence of stimuli. Setting $\lambda = 1/n$ means that λ will approach zero quickly as the number of trials increases, resulting in representations that no longer change. If the bandwidth parameter h is small, and the number of representation points is large, the regular theorems for kernel density estimators ensure that the density estimates will converge to the true stimulus magnitude densities and, hence, the classic SDT model.

Evaluation of the Model

Our initial simulations showed that our model can reproduce classic patterns of data observed in discrimination tasks, including changes in estimates of d' with changes in stimulus discriminability, and changes in estimates of β with changes in signal frequency and payoffs. We then conducted four experiments designed to provide data to which our model could be fit. For each experiment, data from individual subjects were simulated by the model without varying parameters across subjects or across experimental conditions, although the different experiments required changes in the parameters. Table 2 provides a list of the parameters used for the model for each experiment.

For each experiment, we presented the model with the stimulus sequences experienced by each subject and recorded its responses. We estimated traditional SDT statistics (d' and β) and showed that the simulated data matched the human data quite well.

In Experiment 1, when moving from easily discriminable to less discriminable stimuli, the estimates of d' computed from the model's responses decreased while β remained constant, mirroring the changes in the human data. Experiment 2 introduced fluctuations in the frequency of signals over trials. Our subjects were quick to perceive these fluctuations, but there was some delay, as shown in the posterior mean change point distributions. Most importantly, when experiencing the same fluctuations, the model showed the same delays. This effect was not the result of modifications to the decision rule or changes in the model parameters between conditions. Instead, increases (and decreases) in the proportion of "signal" responses over trials arose from a change in the stimulus representations, directly influenced by the stimulus stream. The attention parameter λ and $N_{\max}$ increased to fit the

[8] The kernel estimator ($\lambda = 1/n$) will converge "in rth order mean" to the true density.

data, but the other parameters were the same as those used to fit the data from Experiment 1.

Experiment 3 presented subjects with stimuli drawn from distributions that did not have a monotonic likelihood ratio. Subjects were required to respond "yes" to extreme stimuli in either direction and "no" to stimuli close to the middle of the decision axis. As in Experiment 2, model simulations produced data that closely followed the behavior of the subjects as the probability of a signal changed. Fitting the model required us to reduce the bandwidth parameter *h* relative to Experiment 2. This ensured that the representation of the much narrower noise distribution remained accurate over changes in signal frequency.

In Experiment 4, we explored the behavior of the model when no feedback was provided. In this situation, the model should not know which representation to update after a response. We used an algorithm in which the model assumed that its response was correct, so it presented itself with occasionally inaccurate feedback. The properties of the stimulus stream, the means of the signal and noise distributions, then changed drastically and without warning at the midpoint of the experiment. Like the subjects, the model adjusted relatively quickly to this change if provided with corrective feedback. When no feedback was given, both the model and the subjects, although they changed their effective criteria, could not achieve the same level of performance as the feedback group.

In a further exploration of the role of feedback, we fit the model to data from a recognition memory experiment (Brown et al., 2007). In this experiment, the mirror effect was attenuated after a change in the discriminability of the distractors. Without feedback, our model was able to reproduce both the mirror effect in the early blocks of the experiment and the attenuation and reinstatement of the mirror effect in the later blocks of the experiment after the stimulus change. We also considered data from Mickes et al. (2011) that showed how stronger memories tend to produce extreme confidence ratings, and showed how this finding could result from the sparse representation of the decision axis. Finally, we discussed data from Cox and Dobbins (2011) and emphasized the role of the prior in recognition memory paradigms in which the test list consists of all old or new words.

In all four experiments we simulated the model as if all participants were identical—with identical parameters, differing only in the random selection of stimulus magnitudes they encountered. We took the same approach for the recognition memory data, and even though we were overly constrained, we were able to reproduce changes in the classic mirror effect and the lack of scaling in strong memories. We were able to fit these data without feedback, relying on the naive assumption that observers use their own response to determine which representation to update. To produce the model fits, we adjusted only four parameters: the attention weight (λ), the bandwidth (h), the maximum number of representation points ($N_{\max}$), and the probability of replacing a representation point (γ). It is quite likely that better fits would result from allowing individual participants to have different parameter values.

Extensions to the Model

Although we are generally satisfied with the performance of the dynamic model as an explanation for changes in discrimination performance with changes in the stimulus environment, there are at least two weaknesses that could be addressed.

First, the model deals only with two-choice decisions about univariate stimuli. Extending the model to multidimensional stimuli such as those used in categorization experiments (e.g., Ashby & Gott, 1988; Nosofsky, 1985) is straightforward and requires only that we replace the representation points x_r by vectors indicating points in *n*-dimensional space. The updating kernel would then use, rather than a rectangle of width $2h$, a sphere of radius h to determine those representation points influenced by any particular stimulus. Extending the model to *n*-choice decisions is more complicated. Because the choice in the dynamic model is determined by an evaluation of the (unidimensional) likelihood ratio, we would be forced to consider alternative decision rules. In this situation, we would be dealing with explanations of categorization performance, and there are other, potentially better models designed to accommodate such data (see below).

A second major shortcoming of our model, and of the SDT framework in general, is that it fails to make predictions about response times. Response latency is an important dependent variable that can sometimes adjudicate between theories that make indistinguishable predictions about response choices (see, e.g., Donkin, Brown, Heathcote, & Marley, 2009; Ratcliff & Smith, 2004; Van Zandt, Colonius, & Proctor, 2000). Ratcliff (1978) provided an example of how the traditional SDT framework could be extended through sequential sampling to provide a model for response time as well as choice. The same principle could be applied to our model, building a diffusion model (or some other sampling model) in which the distributions of drift arise from the stimulus representations constructed by the adaptive kernel method.

Implementation of a response latency mechanism is beyond the scope of this article, but it is important to recognize that there are modeling approaches that incorporate response latency mechanisms with proposals about how stimulus representations are established and evolve over time. These include not only the model proposed by Lee and Dry (2006), discussed above, but other work in categorization (e.g., Nosofsky & Palmeri, 1997), attention (e.g., Logan, 2002), and dynamic decision making (e.g., Newell & Lee, 2010).

Relation to Other Models

The discrimination tasks accommodated by SDT are reduced versions of categorization tasks. There are several well-developed models of categorization—reviewed above—that have taken up the challenge of how people learn categories and how they learn to identify stimuli as belonging to each category. These models include Ashby and colleagues' decision bound model (e.g., Ashby, 1992; Ashby & Gott, 1988; Ashby & Maddox, 1992), Nosofsky and colleagues' generalized context model (McKinley & Nosofsky, 1995; Nosofsky, 1986, 1991), and Krusche's ALCOVE model (Kruschke, 1992). It is important to outline the similarities and differences between our model and these other well-established models. We focus extensively on the GCM, because that model is most similar to ours.

The DBM and GCM both assume that categories are represented as locations in a multidimensional psychological space. In GCM, the space stores exemplars, or instances of stimuli that have been presented as members of each category. DBM is less specific about how the members of each category are established but postulates that they exist in the form of distributions of random perceptual information. The major difference between GCM and DBM is in how decisions are made. DBM uses a deterministic decision rule, assigning responses to different locations in the psychological space. If categories differ on only a single dimension, the DBM is equivalent to the SDT model: Observers always respond "signal" when a stimulus is perceived to be more intense than a fixed criterion and "noise" otherwise. More generally, the DBM could also use multiple criteria, but the response would always be determined by where a stimulus was perceived to be located along the decision axis.

The GCM model, by contrast, uses a probabilistic decision rule. A stimulus is compared to each stored exemplar, and a similarity is computed based on the distance between the stimulus and the exemplar. These similarities are summed over all exemplars in each category, and then the probability of responding that the stimulus belongs to a particular category is a function of the summed similarities to exemplars from that category divided by the summed similarities to all exemplars. Under certain careful choices about each model, the GCM and our model share a number of central elements (see, e.g., Ashby & Alfonso-Reese, 1995).

The differences between GCM and our model lie primarily in how points in perceptual space are identified, the information stored in the space, and how decisions are made. Like the DBM, decisions in our model are based on a simple deterministic rule, rather than the probabilistic choice rule of GCM. However, Nosofsky has also explored deterministic versions of the GCM (McKinley & Nosofsky, 1995; Nosofsky, 1991), and Ashby and Maddox (1993) have delineated the conditions under which (in the DBM and GCM) the probabilistic and deterministic choice rules yield equivalent predictions.

Considering the kind of information stored in the models, the comparison between a new stimulus and the representations is more limited in our model than in the GCM. In the GCM, a new stimulus is compared to all the exemplars in each category separately. In our model, the stimulus is compared only to its nearest neighbor. Although this neighbor contains information about the other stimuli that have been presented in that region, it does not contain information about stimuli presented outside the bandwidth. Thus, the likelihood information is not as detailed as the information stored by the GCM. If we were to use a double-exponential kernel, so that even very distant representation points contributed at least some small amount to the likelihood at a particular location, the likelihoods at each representation would be similar to the summed similarities of the GCM model with memory decay and an exponential distance function.

The representation points in our model can be seen as exemplars, but only when there is no updating of the likelihoods at each representation point ($\lambda = 0$) and no memory loss ($\gamma = 0$). This is because representation points are not uniquely identified as either signal or noise unless the updating is eliminated, which in turn eliminates the possibility that at one time the likelihoods favor one response but that at a later time those likelihoods are reversed. Our representation points can move within the perceptual space as new stimuli are presented and old points drop out. Observers can store only a few of these representation points, and information at all other locations in the space is coarsely approximated from these few points. The representation points might not be at any location observed as a value in the stimulus stream, depending on the prior. This cannot happen in the GCM, as all stored exemplars arise from observation, although Nosofsky and colleagues (McKinley & Nosofsky, 1995; Nosofsky, Kruschke, & McKinley, 1992) have explored a version of the GCM model with memory loss, in which recent exemplars are more heavily weighted in the similarity computation than older ones.

The close relationship between GCM and our model suggests that we can, under some circumstances, think of our model as a reduced version of the GCM. The limitations of our model involve mostly a reduced amount of information about the perceptual space and the likelihoods. Unlike the GCM, our model emphasizes dynamic environments. It focuses on building a stimulus representation and then allowing that representation to change with changes to the stimulus distributions. This is the rationale behind our updating process. When the signal likelihoods, for example, are updated, the signal likelihoods of representation points near the observation are increased, whereas the noise likelihoods are decreased. In this sense, the representation points actually compete with one another in a way much like the category nodes in ALCOVE (Kruschke, 1992).

The GCM has proven extraordinarily successful as a model of categorization, and our model's relationship to the GCM provides some assurance that the model might accommodate basic results from many dozens of different categorization experiments. The limitations imposed on our model—most notably the finite number of representation points—mean that it will never be quite equivalent to the GCM, so its ability to model categorization data like the GCM must be confirmed through simulation.

Conclusions

The purpose of our article was not to provide a competitor for the theory of signal detection as it is applied to estimate discriminability and response bias in two-choice tasks. We do not even argue against its use as a metaphor in models of cognition, although we do argue that it provides no explanation of the decision process. We have pointed out (not for the first time) the shortcomings of SDT as a cognitive process model, and proposed an alternative model that does not require a priori stimulus representations and the existence of fixed criteria.

The important contribution of our model is that it provides an explanation for changes in discrimination performance with changes in the stimulus environment. These changes arise from a constantly evolving stimulus representation sensitive to the statistical properties of the stimulus stream.

Our model may not suggest changes in the way theorists use SDT in cognitive models so much as changes in the way they consider the parameters and representations implied by classic SDT. The results of our experiments and simulations suggest that our model can explain not only classical results but also behavior in dynamic decision-making environments and environments with novel decision rules.

References

Ashby, F. G. (Ed.). (1992). *Multidimensional models of perception and cognition.* Hillsdale, NJ: Erlbaum.

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology, 39,* 216–233. doi:10.1006/jmps.1995.1021

Ashby, F. G., & Ell, S. W. (2001). The neurobiology of human category learning. *Trends in Cognitive Sciences, 5,* 204–210. doi:10.1016/S1364-6613(00)01624-7

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 33–53. doi:10.1037/0278-7393.14.1.33

Ashby, F. G., & Maddox, W. T. (1992). Complex decision rules in categorization: Contrasting novice and experienced performance. *Journal of Experimental Psychology: Human Perception and Performance, 18,* 50–71. doi:10.1037/0096-1523.18.1.50

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar and decision bound models of categorization. *Journal of Mathematical Psychology, 37,* 372–400. doi:10.1006/jmps.1993.1023

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review, 93,* 154–179. doi:10.1037/0033-295X.93.2.154

Atkinson, R. C., & Kinchla, R. A. (1965). A learning model for forced-choice detection experiments. *British Journal of Mathematical and Statistical Psychology, 18,* 183–206.

Balakrishnan, J. D. (1998). Some more sensitive measures of sensitivity and response bias. *Psychological Methods, 3,* 68–90. doi:10.1037/1082-989X.3.1.68

Balakrishnan, J. D. (1999). Decision processes in discrimination: Fundamental misrepresentations of signal detection theory. *Journal of Experimental Psychology: Human Perception and Performance, 25,* 1189–1206. doi:10.1037/0096-1523.25.5.1189

Beck, J. R., & Shultz, E. K. (1986). The use of relative operating characteristic (ROC) curves in test performance evaluation. *Archives of Pathology Laboratory Medicine, 110,* 13–20.

Benjamin, A. S., Diaz, M., & Wee, S. (2009). Signal detection with criterion noise: Applications to recognition memory. *Psychological Review, 116,* 84–115. doi:10.1037/a0014351

Bernbach, H. A. (1967). Decision processes in memory. *Psychological Review, 74,* 462–480. doi:10.1037/h0025132

Bernbach, H. A. (1971). Strength theory and confidence ratings in recall. *Psychological Review, 78,* 338–340. doi:10.1037/h0031034

Brenner, L. A. (2003). A random support model of the calibration of subjective probabilities. *Organizational Behavior and Human Decision Processes, 90,* 87–110. doi:10.1016/S0749-5978(03)00004-9

Brown, S. D., Marley, A. A. J., Donkin, C., & Heathcote, A. (2008). An integrated model of choices and response times in absolute identification. *Psychological Review, 115,* 396–425. doi:10.1037/0033-295X.115.2.396

Brown, S., & Steyvers, M. (2005). The dynamics of experimentally induced criterion shifts. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 587–599. doi:10.1037/0278-7393.31.4.587

Brown, S., Steyvers, M., & Hemmer, P. (2007). Modeling experimentally induced strategy shifts. *Psychological Science, 20,* 40–45. doi:10.1111/j.1467-9280.2007.01846.x

Cohen, A. L., Rotello, C. M., & Macmillan, N. A. (2008). Evaluating models of remember–know judgments: Complexity, mimicry, and discriminability. *Psychonomic Bulletin & Review, 15,* 906–926. doi:10.3758/PBR.15.5.906

Cox, J. C., & Dobbins, I. G. (2011). The striking similarities between standard, distractor-free, and target-free recognition. *Memory & Cognition, 39,* 925–940. doi:10.3758/s13421-011-0090-3

Dawes, R. M. (1980). Confidence in intellectual 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: Huber.

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review, 108,* 452–478. doi:10.1037/0033-295X.108.2.452

Donkin, C., Brown, S. D., Heathcote, A., & Marley, A. A. J. (2009). Dissociating speed and accuracy in absolute identification: The effect of unequal stimulus spacing. *Psychological Research, 73,* 308–316. doi:10.1007/s00426-008-0158-2

Dorfman, D. D., & Biderman, M. (1971). A learning model for a continuum of sensory states. *Journal of Mathematical Psychology, 8,* 264–284. doi:10.1016/0022-2496(71)90017-4

Dorfman, D. D., Saslow, C. F., & Simpson, J. C. (1975). Learning models for a continuum of sensory states reexamined. *Journal of Mathematical Psychology, 12,* 178–211. doi:10.1016/0022-2496(75)90056-5

Dunning, D., Heath, C., & Sols, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest, 5,* 69–106. doi:10.1111/j.1529-1006.2004.00018.x

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

Erev, I. (1998). Signal detection by human observers: A cutoff reinforcement learning model of categorization decisions under uncertainty. *Psychological Review, 105,* 280–298. doi:10.1037/0033-295X.105.2.280

Fernberger, S. W. (1920). Interdependence of judgments within the series for the method of constant stimuli. *Journal of Experimental Psychology, 3,* 126–150. doi:10.1037/h0065212

Ferrell, W. R., & McGoey, P. J. (1980). A model of calibration for subjective probabilities. *Organizational Behavior and Human Performance, 26,* 32–53. doi:10.1016/0030-5073(80)90045-8

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91,* 1–67. doi:10.1037/0033-295X.91.1.1

Glanzer, M., Adams, J. K., Iverson, G. J., & Kim, K. (1993). The regularities of recognition memory. *Psychological Review, 100,* 546–567. doi:10.1037/0033-295X.100.3.546

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

Healy, A. F., & Kubovy, M. (1981). Probability matching and the formation of conservative decision rules in a numerical analog of signal detection. *Journal of Experimental Psychology: Human Learning and Memory, 7,* 344–354. doi:10.1037/0278-7393.7.5.344

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 95,* 528–551. doi:10.1037/0033-295X.95.4.528

Hoffmann, C., Pizlo, Z., Popescu, V., & Price, S. (2007). Perception of surfaces from line drawings. *Displays, 28,* 1–7. doi:10.1016/j.displa.2006.11.001

Howarth, C. I., & Bulmer, M. G. (1956). Non-random sequences in visual threshold experiments. *Quarterly Journal of Experimental Psychology, 8,* 163–171. doi:10.1080/17470215608416816

Kac, M. (1962). A note on learning signal detection. *IRE Transactions on Information Theory, 8,* 126–128.

Kac, M. (1969). Some mathematical models in science. *Science, 166,* 695–699. doi:10.1126/science.166.3906.695

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99,* 22–44. doi:10.1037/0033-295X.99.1.22

Kubovy, M., & Healy, A. F. (1977). The decision rule in probabilistic categorization: What it is and how it is learned. *Journal of Experimental Psychology: General, 106,* 427–446. doi:10.1037/0096-3445.106.4.427

Lee, M. D., & Dry, M. J. (2006). Decision making and confidence given uncertain advice. *Cognitive Science, 30,* 1081–1095. doi:10.1207/s15516709cog0000_71

Lee, M. D., & Wagenmakers, E.-J. (2010). *A course in Bayesian* *graphical modeling for cognitive science*. Retrieved from http://www.ejwagenmakers.com/BayesCourse/BayesBookWeb.pdf

Lichtenstein, S., & Fischoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance, 20,* 159–183. doi:10.1016/0030-5073(77)90001-0

Logan, G. D. (2002). An instance theory of attention and memory. *Psychological Review, 109,* 376–400. doi:10.1037/0033-295X.109.2.376

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis.* New York, NY: Wiley.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide.* Mahwah, NJ: Erlbaum.

Maddox, W. T. (2002). Toward a unified theory of decision criterion learning in perceptual categorization. *Journal of the Experimental Analysis of Behavior, 78,* 567–595. doi:10.1901/jeab.2002.78-567

McKinley, S. C., & Nosofsky, R. M. (1995). Investigations of exemplar and decision bound models in large, ill-defined category structures. *Journal of Experimental Psychology: Human Perception and Performance, 21,* 128–148. doi:10.1037/0096-1523.21.1.128

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. doi:10.1037/0096-3445.135.3.391

Metz, C. E., Herman, B. A., & Shen, J. H. (1998). Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. *Statistics in Medicine, 17,* 1033–1053. doi:10.1002/(SICI)1097-0258(19980515)17:9<1033::AID-SIM784>3.0.CO;2-Z

Mickes, L., Hwe, V., Wais, P. E., & Wixted, J. T. (2011). Strong memories are hard to scale. *Journal of Experimental Psychology: General, 140,* 239–257. doi:10.1037/a0023007

Mueller, S. T., & Weidemann, C. T. (2008). Decision noise: An explanation for observed violations of signal detection theory. *Psychonomic Bulletin & Review, 15,* 465–494. doi:10.3758/PBR.15.3.465

Murdock, B. B., Jr. (1965). Signal-detection theory and short-term memory. *Journal of Experimental Psychology, 70,* 443–447. doi:10.1037/h0022543

Murdock, B. B., Jr. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review, 89,* 609–626. doi:10.1037/0033-295X.89.6.609

Murdock, B. B., Jr., & Dufty, P. O. (1972). Strength theory and recognition memory. *Journal of Experimental Psychology, 94,* 284–290. doi:10.1037/h0032795

Newell, B. R., & Lee, M. D. (2010). The right tool for the job? Comparing an evidence accumulation and a naive strategy selection model of decision making. *Journal of Behavioral Decision Making.* Advance online publication. doi:10.1002/bdm.703

Nosofsky, R. M. (1983). Information integration and the identification of stimulus noise and criterial noise in absolute judgment. *Journal of Experimental Psychology: Human Perception and Performance, 9,* 299–309. doi:10.1037/0096-1523.9.2.299

Nosofsky, R. M. (1985). Overall similarity and the identification of separable-dimension stimuli: A choice model analysis. *Perception & Psychophysics, 38,* 415–432. doi:10.3758/BF03207172

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General, 115,* 39–57. doi:10.1037/0096-3445.115.1.39

Nosofsky, R. M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception and Performance, 17,* 3–27. doi:10.1037/0096-1523.17.1.3

Nosofsky, R. M., Kruschke, J. K., & McKinley, S. C. (1992). Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 211–233. doi:10.1037/0278-7393.18.2.211

Nosofsky, R. M., & Palmeri, T. J. (1997). Comparing exemplar-retrieval and decision-bound models of speeded perceptual classification. *Perception & Psychophysics, 59,* 1027–1048. doi:10.3758/BF03205518

Nosofsky, R. M., & Stanton, D. (2005). Speeded classification in a probabilistic category structure: Contrasting exemplar-retrieval, decision-boundary, and prototype models. *Journal of Experimental Psychology: Human Perception and Performance, 31,* 608–629. doi:10.1037/0096-1523.31.3.608

Oja, E. (1982). A simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology, 15,* 267–273. doi:10.1007/BF00275687

Petrov, A., & Anderson, J. R. (2005). The dynamics of scaling: A memory-based anchor model of category rating and absolute identification. *Psychological Review, 112,* 383–416. doi:10.1037/0033-295X.112.2.383

Pleskac, T. J., & Busemeyer, J. R. (2010). *Two-stage dynamic signal detection theory: A dynamic and stochastic theory of confidence, choice, and response time.* Manuscript submitted for publication.

Rabbitt, P. M. A. (1981). Sequential reactions. In D. H. Holding (Ed.), *Human skills* (pp. 153–175). London, England: Wiley.

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

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science, 9,* 347–356. doi:10.1111/1467-9280.00067

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

Ratcliff, R., & Starns, J. J. (2009). Modeling confidence and response time in recognition memory. *Psychological Review, 116,* 59–83. doi:10.1037/a0014086

Rotello, C. M., Macmillan, N. A., & Reeder, J. A. (2004). Sum–difference theory of remembering and knowing: A two-dimensional signal-detection model. *Psychological Review, 111,* 588–616. doi:10.1037/0033-295X.111.3.588

Rouder, J. N., & Ratcliff, R. (2004). Comparing categorization models. *Journal of Experimental Psychology: General, 133,* 63–82. doi:10.1037/0096-3445.133.1.63

Schnyer, D. M., Maddox, W. T., Ell, S., Davis, S., Pacheco, J., & Verfaellie, M. (2009). Prefrontal contributions to rule-based and information-integration category learning. *Neuropsychologia, 47,* 2995–3006. doi:10.1016/j.neuropsychologia.2009.07.011

Schoeffler, M. S. (1965). Theory for psychophysical learning. *Journal of the Acoustical Society of America, 37,* 1124–1133. doi:10.1121/1.1909534

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—Retrieving effectively from memory. *Psychonomic Bulletin & Review, 4,* 145–166. doi:10.3758/BF03209391

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London, England: Chapman & Hall.

Slotnick, S. D. (2010). "Remember" source memory ROCs indicate recollection is a continuous process. *Memory, 18,* 27–39. doi:10.1080/09658210903390061

Smith, A. F. M. (1975). A Bayesian approach to inference about a change-point in a sequence of random variables. *Biometrika, 62,* 407–416. doi:10.1093/biomet/62.2.407

Strong, M. H., & Strong, E. K., Jr. (1916). The nature of recognition memory and of the localization of recognition. *American Journal of Psychology, 27,* 341–362. doi:10.2307/1413103

Swets, J. A., Tanner, W. P., Jr., & Birdsall, T. G. (1961). Decision processes in perception. *Psychological Review, 68,* 301–340. doi:10.1037/h0040547

Thomas, E. A. (1973). On a class of additive learning models: Error-correcting and probability matching. *Journal of Mathematical Psychology, 10,* 241–264. doi:10.1016/0022-2496(73)90017-5

Treisman, M., & Williams, T. C. (1984). A theory of criterion setting with an application to sequential dependencies. *Psychological Review, 91,* 68–111. doi:10.1037/0033-295X.91.1.68

Underwood, B. J. (1972). Word recognition memory and frequency information. *Journal of Experimental Psychology, 94,* 276–283. doi:10.1037/h0032785

Van Es, J. J., Vladusich, T., & Cornelissen, F. W. (2007). Local and relational judgements of surface colour: Constancy indices and discrimination performance. *Spatial Vision, 20,* 139–154. doi:10.1163/156856807779369733

Van Zandt, T. (2000). How to fit a response time distribution. *Psychonomic Bulletin & Review, 7,* 424–465. doi:10.3758/BF03214357

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. doi:10.3758/BF03212980

Van Zandt, T., & Jones, M. R. (2011). *Stimulus rhythm and choice performance.* Manuscript submitted for publication.

Verplanck, W. S., Collier, G. H., & Cotton, J. W. (1952). Nonindependence of successive responses in measurements of the visual threshold. *Journal of Experimental Psychology, 44,* 273–282. doi:10.1037/h0054948

Verplanck, W. S., & Cotton, J. W. (1955). The dependence of frequencies of seeing on procedural variables: I. Direction and length of series of intensity-ordered stimuli. *Journal of General Psychology, 53,* 37–47. doi:10.1080/00221309.1955.9710135

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

Vickers, D. (1982). Effects of alternating set for speech or accuracy on response time, accuracy and confidence in a unidimensional discrimination task. *Acta Psychologica, 50,* 179–197. doi:10.1016/0001-6918(82)90006-3

Vickers, D., & Lee, M. D. (1998). Dynamic models of simple judgments: I. Properties of a self-regulating accumulator module. *Nonlinear Dynamics, Psychology, and Life Sciences, 2,* 169–194. doi:10.1023/A:1022371901259

Vickers, D., & Lee, M. D. (2000). Dynamic models of simple judgments: II. Properties of a self-organizing PAGAN (Parallel, Adaptive, Generalized Accumulator Network) model for multi-choice tasks. *Nonlinear Dynamics, Psychology, and Life Sciences, 4,* 1–31. doi:10.1023/A:1009571011764

Wallace, W. P. (1978). Recognition failure of recallable words and recognizable words. *Journal of Experimental Psychology: Human Learning and Memory, 4,* 441–452. doi:10.1037/0278-7393.4.5.441

Wallace, W. P. (1982). Distractor-free recognition tests of memory. *American Journal of Psychology, 95,* 421–440. doi:10.2307/1422134

Wallace, W. P., Sawyer, T. J., & Robertson, L. C. (1978). Distractors in recall, distractor-free recognition, and the word-frequency effect. *American Journal of Psychology, 91,* 295–304. doi:10.2307/1421539

Wallsten, T. S., & González-Vallejo, C. (1994). Statement verification: A stochastic model of judgment and response. *Psychological Review, 101,* 490–504. doi:10.1037/0033-295X.101.3.490

Wickelgren, W. A., & Norman, D. A. (1966). Strength models and serial position in short-term recognition memory. *Journal of Mathematical Psychology, 3,* 316–347. doi:10.1016/0022-2496(66)90018-6

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review, 114,* 152–176. doi:10.1037/0033-295X.114.1.152

Wixted, J. T., & Stretch, V. (2004). In defense of the signal detection interpretation of remember/know judgments. *Psychonomic Bulletin & Review, 11,* 616–641. doi:10.3758/BF03196616

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. doi:10.1037/0278-7393.20.6.1341

Yonelinas, A. P. (1999). The contribution of recollection and familiarity to recognition and source-memory judgments: A formal dual-process model and an analysis of receiver operating characteristics. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 1415–1434. doi:10.1037/0278-7393.25.6.1415

Received April 28, 2010
Revision received June 14, 2011
Accepted June 24, 2011 ■