1-2009

Signal detection with criterion noise: Applications to recognition memory

Aaron S. BENJAMIN

Michael DIAZ

WEE, Serena

Singapore Management University, serenawee@smu.edu.sg

DOI: https://doi.org/10.1037/a0014351

Signal Detection With Criterion Noise: Applications to Recognition Memory

Aaron S. Benjamin, Michael Diaz, and Serena Wee
University of Illinois at Urbana–Champaign

A tacit but fundamental assumption of the theory of signal detection is that criterion placement is a noise-free process. This article challenges that assumption on theoretical and empirical grounds and presents the noisy decision theory of signal detection (ND-TSD). Generalized equations for the isosensitivity function and for measures of discrimination incorporating criterion variability are derived, and the model's relationship with extant models of decision making in discrimination tasks is examined. An experiment evaluating recognition memory for ensembles of word stimuli revealed that criterion noise is not trivial in magnitude and contributes substantially to variance in the slope of the isosensitivity function. The authors discuss how ND-TSD can help explain a number of current and historical puzzles in recognition memory, including the inconsistent relationship between manipulations of learning and the isosensitivity function's slope, the lack of invariance of the slope with manipulations of bias or payoffs, the effects of aging on the decision-making process in recognition, and the nature of responding in remember–know decision tasks. ND-TSD poses novel, theoretically meaningful constraints on theories of recognition and decision making more generally, and provides a mechanism for rapprochement between theories of decision making that employ deterministic response rules and those that postulate probabilistic response rules.

Keywords: signal detection, recognition memory, criteria, decision making

The theory of signal detection (TSD[1]; Green & Swets, 1966; Macmillan & Creelman, 2005; Peterson, Birdsall, & Fox, 1954; Tanner & Swets, 1954) is a theory of decision making that has been widely applied to psychological tasks involving detection, discrimination, identification, and choice, as well as to problems in engineering and control systems. Its historical development follows quite naturally from earlier theories in psychophysics (Blackwell, 1953; Fechner, 1860; Thurstone, 1927) and advances in statistics (Wald, 1950). The general framework has proven sufficiently flexible so as to allow substantive cross-fertilization with related areas in statistics and psychology, including mixture distributions (DeCarlo, 2002), theories of information integration in multidimensional spaces (Banks, 2000; Townsend & Ashby, 1982), models of group decision making (Sorkin & Dai, 1994), models of response timing (D. A. Norman & Wickelgren, 1969; Sekuler, 1965; Thomas & Myers, 1972), and multiprocess models that combine thresholded and continuous evidence distributions (Yonelinas, 1999). It also exhibits well-characterized relationships with other prominent perspectives, such as individual choice theory (Luce, 1959) and threshold-based models (Krantz, 1969; Swets, 1986b). Indeed, it is arguably the most widely used and successful theoretical framework in psychology of the past half century.

The theoretical underpinnings of TSD can be summarized in four basic postulates:

1. Events are individual enumerable *trials* on which a signal is present or not.
2. A *strength* value characterizes the evidence for the presence of the signal on a given trial.
3. Random variables characterize the conditional *probability distributions* of strength values for signal-present and signal-absent events (for detection) or for Signal A and Signal B events (for discrimination).
4. A *criterion* serves to map the continuous strength variable (or its associated likelihood ratio) onto a binary (or *n*-ary) decision variable.

As applied to recognition memory experiments (Banks, 1970; Egan, 1958; Lockhart & Murdock, 1970; Parks, 1966), in which subjects make individual judgments about whether a test item was

Aaron S. Benjamin, Michael Diaz, and Serena Wee, Department of Psychology, University of Illinois at Urbana–Champaign.

This work was supported by National Institutes of Health Grant R01 AG026263. We offer many thanks for useful commentary and suggestions to the Human Memory and Cognition Lab at the University of Illinois at Urbana–Champaign (http://www.psych.uiuc.edu/~asbenjam/), John Wixted, and Gary Dell.

Correspondence concerning this article should be addressed to Aaron S. Benjamin, Department of Psychology, University of Illinois, Urbana–Champaign, Champaign, IL 61820. E-mail: asbenjam@uiuc.edu

[1] More commonly, this theory is referred to as *signal-detection theory* (e.g., Swets, 1964). Here, the alternative acronym TSD is preferred (see also Birdsall, 1966; Lockhart & Murdock, 1970; Tanner, 1960) in that it properly emphasizes the theory's relation to, but not isomorphism with, statistical decision theory (Wald, 1950).

previously viewed in a particular delimited study episode, the signal is considered to be the prior study of the item. That study event is thought to confer additional strength on the item such that studied items generally, but not always, yield greater evidence for prior study than do unstudied items. Subjects then make a decision about whether they did or did not study the item by comparing the strength yielded by the current test stimulus to a decision criterion. Analytically, TSD reparameterizes the obtained experimental statistics as estimates of discriminability and response criterion or bias. Theoretical conclusions about the mnemonic aspects of recognition performance are often drawn from the form of the *isosensitivity function*,[2] which is a plot of the theoretical hit rate against the theoretical false-alarm rate across all possible criterion values. The function is typically estimated from points derived from a confidence-rating procedure (Egan, 1958; Egan, Schulman, & Greenberg, 1959).

TSD has been successfully applied to recognition because it provides an articulated and intuitive description of the decision portion of the task without obliging any particular theoretical account of the relevant memory processes. In fact, theoretical interpretations derived from the application of TSD to recognition memory have been cited as major constraints on process models of recognition (e.g., McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997). Recent evidence reveals an increased role of TSD in research on recognition memory: The number of citations in PsycINFO that appear in response to a joint query of *recognition memory* and *signal detection* as keywords has increased from 23 in the 1980s to 39 in the 1990s to 67 in just the first seven years of this decade.

The purpose of this article is to theoretically and empirically evaluate the postulate of a noise-free criterion (Assumption 4, above) and to describe an extension of TSD that is sufficiently flexible to handle criterion variability. The claim is that criteria may vary from trial to trial in part because of noise inherent in the processes involved with maintaining and updating them. Although this claim does not seriously violate the theoretical structure of TSD, it does have major implications for how we draw theoretical conclusions about memory, perception, and decision processes from detection, discrimination, and recognition experiments.

As we review below, concerns about variability in the decision process are apparent in a variety of literatures, and theoretical tools have been advanced to address the problems that arise from noisy decision making (Rosner & Kochaniski, 2008). However, theorizing in recognition memory has mostly advanced independently of such concerns, perhaps in part because of the difficulty associated with disentangling decision noise from representational noise (see, e.g., Ratcliff & Starns, in press). This article considers the statistical and analytic problems that arise from its postulation in the context of detection theoretical models and applies a novel experimental task—the ensemble recognition paradigm—toward the problem of estimating criterion variance.

Historical Antecedents and Contemporary Motivation

Considerations similar to the ones forwarded here have been previously raised in the domains of psychoacoustics and psychophysics (Durlach & Braida, 1969; Gravetter & Lockhead, 1973) but have not been broadly considered in the domain of recognition memory. An exception is the seminal "strength theory" of Wickelgren (1968; Wickelgren & Norman, 1966; D. A. Norman & Wickelgren, 1969), on whose work our initial theoretical rationale is based. That work was applied predominately to problems in short-term memory and to the question of how absolute (yes–no) and relative (forced-choice) response tasks differ from one another. However, general analytic forms for the computation of detection statistics were not provided, nor was the work applied to the relationship between the isosensitivity function and theories of recognition memory (which were not prominent at the time).

Contemporary versions of the TSD are best understood by their relation to the general class of judgment models derived from Thurstone (1927). A taxonomy of those models described by Torgerson (1958) allows various restrictions on the equality of stimulus variance and of criterial variance; current applications of TSD to recognition memory vary in whether they permit stimulus variance to differ across distributions, but they almost unilaterally disallow criterial variance. This is a restriction that, although not unique to this field, is certainly a surprising dissimilarity with work in related areas such as detection and discrimination in psychophysical tasks (Bonnel & Miller, 1994; Durlach & Braida, 1969; Nosofsky, 1983) and classification (Ashby & Maddox, 1993; Erev, 1998; Kornbrot, 1980). The extension of TSD to the noisy decision theory of signal detection (ND-TSD) is a relaxation of this restriction: ND-TSD permits nonzero criterial variance.

The recent explosion of work evaluating the exact form of the isosensitivity function in recognition memory under different conditions (Arndt & Reder, 2002; Glanzer, Kim, Hilford, & Adams, 1999; Gronlund & Elam, 1994; Kelley & Wixted, 2001; Matzen & Benjamin, in press; Qin, Raye, Johnson, & Mitchell, 2001; Ratcliff, McKoon, & Tindall, 1994; Ratcliff, Sheu, & Gronlund, 1992; Slotnick, Klein, Dodson, & Shimamura, 2000; Van Zandt, 2000; Yonelinas, 1994, 1997, 1999) and in different populations (M. R. Healy, Light, & Chung, 2005; Howard, Bessette-Symons, Zhang, & Hoyer, 2006; Manns, Hopkins, Reed, Kitchener, & Squire, 2003; Wixted & Squire, 2004a, 2004b; Yonelinas, Kroll, Dobbins, Lazzara, & Knight, 1998; Yonelinas et al., 2002, 2004), as well as the prominent role those functions play in current theoretical development (Dennis & Humphreys, 2001; Glanzer, Adams, Iverson, & Kim, 1993; McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997; Wixted, 2007; Yonelinas, 1999), suggests the need for a thorough reappraisal of the underlying variables that contribute to those functions. Because work in psychophysics (Krantz, 1969; Nachmias & Steinman, 1963) and, more recently, in recognition memory (Malmberg, 2002; Malmberg & Xu, 2006; Wixted & Stretch, 2004) has illustrated how aspects and suboptimalities of the decision process can influence the shape of the isosensitivity function, the goals of this article are to provide an organizing framework for the incorporation of decision noise within TSD and to help expand the various theoretical discussions within the field of recognition memory to include a role for decision variability.

[2] Following the suggestion of Luce (1963), we use the term *isosensitivity function* instead of the more historically relevant but somewhat unintuitive label of *receiver* (or *relative*) *operating characteristic* (ROC). Throughout this article, no change in terminology is used to indicate whether the isosensitivity function is plotted in probability or normal-deviate coordinates, other than to the relevant axes in figures, unless the distinction is relevant to that discussion.

We suggest that drawing conclusions about the theoretical components of recognition memory from the form of the isosensitivity function can be a dangerous enterprise and show how a number of historical and current puzzles in the literature may benefit from a consideration of criterion noise.

Organization of the Article

The first part of this article provides a short background on the assumptions of traditional TSD models, as well as evidence bearing on the validity of those assumptions. Appreciating the nature of the arguments underlying the currently influential *unequal-variance* version of TSD is critical to understanding the principle of criterial variance and the proposed analytic procedure for separately estimating criterial and evidence variance. In the second part of the article, we critically evaluate the assertion of a stationary and nonvariable scalar criterion value[3] from a theoretical and empirical perspective, and in the third section, we provide basic derivations for the form of the isosensitivity function in the presence of nonzero criterial variability. The fourth portion of the article provides derivations for measures of accuracy in the presence of criterial noise and leads to the presentation, in the fifth section, of the ensemble recognition task, which can be used to assess criterial noise. In the sixth part of the article, different models of that experimental task are considered and evaluated, and estimates of criterial variability are provided. In the seventh and final part of the article, we review the implications of the findings and some of the situations in which a consideration of criterial variability might advance our progress on a number of interesting problems in recognition memory and beyond.

It is important to note that the successes of TSD have led to many unanswered questions and that a reconsideration of basic principles like criterion invariance may provide insight into those problems. No less an authority than John Swets—the researcher most responsible for introducing TSD to psychology—noted that it was unclear why, for example, the slope of the isosensitivity line for detection of brain tumors was approximately half the slope of the isosensitivity line for detection of abnormal tissue cells (Swets, 1986a). Within the field of recognition memory, there is evidence that certain manipulations that lead to increased accuracy, such as increased study time, are also associated with decreased slope of the isosensitivity function (Glanzer et al., 1999; Hirshman & Hostetter, 2000), whereas other manipulations that also lead to superior performance are not (Ratcliff et al., 1994). Although there are extant theories that account for changes in slope, there is no agreed-upon mechanism by which they do so, nor is an explanation of such heterogeneous effects forthcoming.

Throughout this article, we make reference to the recognition decision problem, but most of the considerations presented here are relevant to other problems in detection and discrimination, and we hope that the superficial application to recognition memory will not deter from the more general message about the need to consider decision-based noise in such problems (see also Durlach & Braida, 1969; Gravetter & Lockhead, 1973; Nosofsky, 1983; Wickelgren, 1968).

Assumptions About Evidence Distributions

The lynchpin theoretical apparatus of TSD is the probabilistic relationship between signal status and perceived evidence. The historical assumption about this relationship is that the distributions of the random variables are normal in form (Thurstone, 1927) and of equal variance, separated by some distance, d' (Green & Swets, 1966). Whereas the former assumption has survived inquiry, the latter has been less successful.

The original (Peterson et al., 1954) and most popularly applied version of TSD assumes that signal and noise distributions are of equal variance. Although many memory researchers tacitly endorse this assumption by reporting summary measures of discrimination and criterion placement that derive from the application of the equal-variance model, such as d' and C_j, respectively, the empirical evidence does not support the equal-variance assumption. The slope of the isosensitivity function in recognition memory is often found to be ~0.80 (Ratcliff et al., 1992), although this value may change with increasing discriminability (Glanzer et al., 1999; Heathcote, 2003; Hirshman & Hostetter, 2000). This result has been taken to imply that the evidence distribution for studied items is of greater variance than the distribution for unstudied items (Green & Swets, 1966). The magnitude of this effect, not its existence, and whether manipulations that enhance or attenuate it are actually affecting representational variance, are the issues at stake here.

The remarkable linearity of the isosensitivity function notwithstanding, it is critical for present purposes to note not only that the mean slope for recognition memory is often less than 1 but also that it varies considerably over situations and individuals (Green & Swets, 1966). It is considerably lower than 0.8 for some tasks (e.g., ~0.6 for the detection of brain tumors; Swets et al., 1979), higher than 1 for other tasks (e.g., information retrieval; Swets, 1969), and around 1.0 for yet others (e.g., odor recognition; Rabin & Cain, 1984).

Nonunit and variable slopes reveal an inadequacy of the equal-variance model of Peterson et al. (1954) and of the validity of the measure d'. This failure can be addressed in several ways. It might be assumed that the distributions of evidence are asymmetric in form, for example, or that one or the other distribution reflects a mixture of latent distributions (DeCarlo, 2002; Yonelinas, 1994). The traditional and still predominant explanation, however, is the one described above—that the variance of the distributions is unequal (Green & Swets, 1966; Wixted, 2007). Because the slope of the isosensitivity function is equal to the ratio of the standard deviations of the noise and signal distributions in the unequal-variance TSD model, the empirical estimates of slope less than 1 have promoted the inference that the signal distribution is of greater variance than the noise distribution in recognition. However, the statistical theory of the form of the isosensitivity function that is used to understand nonunit slopes and slope variability has been only partially unified with the psychological theories that produce such behavior, via either the interactivity of continuous and thresholding mechanisms (Yonelinas, 1999) or the averaging

[3] Criterion is used throughout to refer to the location of a decision threshold in the units of the evidence dimension (i.e., in terms of the values on the abscissa in Figure 1, which are typically standard deviations of the noise distribution). *Bias*—a term sometimes used interchangeably with criterion—refers specifically to the value of the likelihood ratio at criterion. In the discussion here, the distinction is often not relevant, in which case we use the term *criterion.*

process presumed by global matching mechanisms (Gillund & Shiffrin, 1984; Hintzman, 1986; Humphreys, Pike, Bain, & Tehan, 1989; Murdock, 1982). None of these prominent theories include a role for criterial variability, nor do they provide a comprehensive account of the shape of isosensitivity functions and of the effect of manipulations on that shape. Criterial variability can directly affect the slope of the isosensitivity function, a datum that opens up novel theoretical possibilities for psychological models of behavior underlying the isosensitivity function.

The form of the isosensitivity function has been used to test the validity of assumptions built into TSD about the nature of the evidence distributions, as well as to estimate parameters for those distributions. In that sense, TSD can be said to have bootstrapped itself into its current position of high esteem: Its validity has mostly been established by confirming its implications, rather than by systematically testing its individual assumptions. This is not intended to be a point of criticism, but it must be kept in mind that the accuracy of such estimation and testing depends fundamentally on the joint assumptions that evidence is inherently variable and that criterion location is not. Allowing criterial noise to play a role raises the possibility that previous explorations of the isosensitivity function in recognition memory have conflated the contributions of stimulus and criterial noise.

Evidence for Criterion Variability

As noted earlier, traditional TSD assumes that criterion placement is a noise-free and stationary process. Although there is some acknowledgment of the processes underlying criterion inconsistency (see, e.g., Macmillan & Creelman, 2005, p. 46), the apparatus of criterion placement in TSD stands in stark contrast to the central assumption of stimulus-related variability (see also Rosner & Kochanski, 2008). There are numerous reasons to doubt the validity of the idea that criteria are noise free. First, there is evidence from detection and discrimination tasks of response autocorrelations, as well as systematic effects of experimental manipulations on response criteria. Second, maintaining the values of one or multiple criteria poses a memory burden and should thus be subject to forgetting and memory distortion. Third, comprehensive models of response time and accuracy in choice tasks suggest the need for criterial variability. Fourth, there is evidence from basic and well-controlled psychophysical tasks of considerable trial-to-trial variability in the placement of criteria. Fifth, there are small but apparent differences between forced-choice response tasks and yes–no response tasks that indicate a violation of one of the most fundamental relationships predicted by TSD: the equality of the area under the isosensitivity function as estimated by the rating procedure and the proportion of correct responses in a two-alternative forced-choice task. This section reviews each of these arguments more fully.

In each case, it is important to distinguish between systematic and nonsystematic sources of variability in criterion placement. This distinction is critical because only nonsystematic variability violates the actual underlying principle of a nonvariable criterion. Some scenarios violate the usual use, but not the underlying principles, of TSD. This section identifies some sources of systematic variability and outlines the theoretical mechanisms that have been invoked to handle them. We also review evidence for nonsystematic sources of variability. Systematic sources of variability can be modeled within TSD by allowing criterion measures to vary with experimental manipulations (Benjamin, 2001; Benjamin & Bawa, 2004; S. Brown & Steyvers, 2005; S. Brown, Steyvers, & Hemmer, 2007), by postulating a time-series criterion localization process contingent upon feedback (Atkinson, Carterette, & Kinchla, 1964; Atkinson & Kinchla, 1965; Friedman, Carterette, Nakatani, & Ahumada, 1968) only following errors (Kac, 1962; Thomas, 1973) or only following correct responses (Model 3 of Dorfman & Biderman, 1971), or as a combination of a long-term learning process and nonrandom momentary fluctuations (Treisman, 1987; Treisman & Williams, 1984). Criterial variance can even be modeled with a probabilistic responding mechanism (Parks, 1966; Thomas, 1975; White & Wixted, 1999), although the inclusion of such a mechanism violates much of the spirit of TSD.

Nonstationarity

When data are averaged across trials to compute TSD parameters, the researcher is tacitly assuming that the criterion is invariant across those trials. By extension, when parameters are computed across an entire experiment, measures of discriminability and criterion are only valid when the criterion is stationary over that entire period. Unfortunately, there is a abundance of evidence that this condition is rarely, if ever, met.

Response autocorrelations. Research more than half a century ago established the presence of longer runs of responses than would be expected under a response-independence assumption (Fernberger, 1920; Howarth & Bulmer, 1956; McGill, 1957; Shipley, 1961; Verplanck, Collier, & Cotton, 1952; Verplanck, Cotton, & Collier, 1953; Wertheimer, 1953). More recently, response autocorrelations (Gilden & Wilson, 1995; Luce, Nosofsky, Green, & Smith, 1982; Staddon, King, & Lockhead, 1980) and response time autocorrelations (Gilden, 1997, 2001; Van Orden, Holden, & Turvey, 2003) within choice tasks have been noted and evaluated in terms of long-range fractal properties (Gilden, 2001; Thornton & Gilden, in press) or short-range response dependencies (Wagenmakers, Farrell, & Ratcliff, 2004, 2005). Such dependencies have even been reported in the context of tasks eliciting confidence ratings (Mueller & Weidemann, 2008). Numerous models were proposed to account for short-range response dependencies, most of which include a mechanism for the adjustment of the response criterion on the basis of feedback of one sort or another (e.g., Kac, 1962; Thomas, 1973; Treisman, 1987; Treisman & Williams, 1984). Because criterion variance was presumed to be systematically related to aspects of the experiment and the subject's performance, however, statistical models that incorporate random criterial noise were not applied to such tasks (e.g., Durlach & Braida, 1969; Gravetter & Lockhead, 1973; Wickelgren, 1968).

The presence of such response correlations in experiments in which the signal value is uncorrelated across trials implies shifts in the decision regime, either in terms of signal reception or transduction or in terms of criterion location. To illustrate this distinction, consider a typical subject in a detection experiment whose interest and attention fluctuate with surrounding conditions (did an attractive research assistant just pass by the door?) and changing internal states (increasing hunger or boredom). If these distractions cause the subject to attend less faithfully to the experiment for a period of time, it could lead to systematically biased evidence

values and, thus, biased responses. Alternatively, if a subject's criterion fluctuates because such distraction affects the subject's ability to maintain a stable value, it will bias responses equivalently from the decision-theoretic perspective. More importantly, fluctuating criteria can lead to response autocorrelations even when the transduction mechanism does not lead to correlated evidence values. Teasing apart these two sources of variability is the major empirical difficulty of our current enterprise.

Effects of experimental manipulations. Stronger evidence for the lability of criteria comes from tasks in which experimental manipulations are shown to induce strategic changes. Subjects appear to modulate their criterion on the basis of their estimated degree of learning (Hirshman, 1995) and perceived difficulty of the distractor set in recognition (Benjamin & Bawa, 2004; S. Brown et al., 2007). Subjects even appear to dynamically shift criteria in response to item characteristics, such as idiosyncratic familiarity (J. Brown, Lewis, & Monk, 1977) and word frequency (Benjamin, 2003). In addition, criteria exhibit reliable individual differences as a function of personality traits (Benjamin, Wee, & Roberts, 2008), thus suggesting another unmodeled source of variability in detection tasks.

It is important to note, however, that criterion changes do not always appear when expected (e.g., Higham, Perfect, & Bruno, in press; Stretch & Wixted, 1998; Verde & Rotello, 2007) and are rarely of an optimal magnitude. It is for this reason that there is some debate over whether subject-controlled criterion movement underlies all of the effects that it has been invoked to explain (Criss, 2006) and indeed, more generally, over whether a reconceptualization of the decision variable itself provides a superior explanation to that of strategic criterion setting (for a review in the context of "mirror effects," see Greene, 2007). For present purposes, it is worth noting that this inconsistency may well reflect the fact that criterion maintenance imposes a nontrivial burden on the rememberer and that he or she may occasionally forgo strategic shifting to minimize the costs of allocating the resources to do so.

These many contributors to criterion variability make it likely that every memory experiment contains a certain amount of systematic but unattributed sources of variance that may affect interpretations of the isosensitivity function if not explicitly modeled. To be clear, such effects are the province of the current model only if they are undetected and unincorporated into the application of TSD to the data. The systematic variability evident in strategic criterion movement may, depending on the nature of that variability, meet the assumptions of ND-TSD and thus be accounted for validly, but we explicitly deal with purely nonsystematic variability in our statistical model.

The Memory Burden of Criterion Maintenance

Given the many systematic sources of variance in criterion placement, it is unlikely that recapitulation of criterion location from trial to trial is a trivial task for the subject. The current criterion location is determined by some complex function relating past experience, implicit and explicit payoffs, and experience thus far in the test, and retrieval of the current value is likely prone to error—a fact that may explain why intervening or unexpected tasks or events that disrupt the normal pace or rhythm of the test appear to affect criterion placement (Hockley & Niewiadomski, 2001). Evidence for this memory burden is apparent when comparing the form of isosensitivity functions estimated from rating procedures with estimates from other procedures, such as payoff manipulations.

Differences between rating-scale and payoff procedures. The difficulty of criterion maintenance is exacerbated in experiments in which confidence ratings are gathered because the subject is forced to maintain multiple criteria, one for each confidence boundary. Although it is unlikely that these values are maintained as independent entities (Stretch & Wixted, 1998), the burden nonetheless increases with the number of required confidence boundaries. Variability introduced by the confidence-rating procedure may explain why the isosensitivity function differs slightly when estimated with that procedure as compared with experiments that manipulate payoff matrices, as well as why rating-derived functions change shape slightly but unexpectedly when the prior odds of signal and noise are varied (Balakrishnan, 1998a; Markowitz & Swets, 1967; Van Zandt, 2000). These findings have been taken to indicate a fundamental failing of the basic assumptions of TSD (Balakrishnan, 1998a, 1998b, 1999) but may simply reflect the contribution of criterion noise (Mueller & Weidemann, 2008).

Deviations of the yes–no decision point on the isosensitivity function. A related piece of evidence comes from the comparison of isosensitivity functions from rating procedures with single points derived from a yes–no judgment. As noted by Wickelgren (1968), it is not uncommon for that yes–no point to lie slightly above the isosensitivity function (Egan, Greenberg, & Schulman, 1961; Markowitz & Swets, 1967; Schulman & Mitchell, 1966; Watson, Rilling, & Bourbon, 1964; Wickelgren & Norman, 1966) and for that effect to be somewhat larger when more confidence categories are employed. This result likely reflects the fact that the maintenance of criteria becomes more difficult with increasing numbers of criterion points. In recognition memory, Benjamin, Lee, and Diaz (2008) showed that discrimination between previously studied and unstudied words was measured to be superior when subjects made yes–no discrimination judgments than when they used a 4-point response scale, as well as superior on the 4-point response scale when compared with an 8-point response scale. This result is consistent with the idea that each criterion introduces noise to the decision process and that, in the traditional analysis, that noise inappropriately contributes to estimates of memory for the studied materials.

Sampling Models of Choice Tasks

A third argument in favor of criterion variance comes from sequential sampling models that explicitly account for both response time and accuracy in two-choice decisions. Specifically, the diffusion model of Ratcliff (1978, 1988; Ratcliff & Rouder, 1998) serves as a benchmark in the field of recognition memory (e.g., Ratcliff, Thapar, & McKoon, 2004) in that it successfully accounts for aspects of data, including response times, that other models do not explicitly address. It would thus seem that general, heuristic models like TSD have much to gain from analyzing the nature of the decision process in the diffusion model.

That model provides a full account of recognition memory only when two critical parameters are allowed to vary (Ratcliff & Rouder, 1998). First is a parameter that corresponds to the variability in the rate with which evidence accumulates from

trial to trial. This value corresponds naturally to stimulus-based variability and resembles the parameter governing variability in the evidence distributions in TSD. The second parameter corresponds to trial-to-trial variability in the starting point for the diffusion process. When this value moves closer to a decision boundary, less evidence is required prior to a decision—thus, this value is analogous to variability in criterion placement. A recent extension of the diffusion model to the confidence-rating procedure (Ratcliff & Starns, in press) has a similar mechanism. The fact that the otherwise quite powerful diffusion model fails to provide a comprehensive account of recognition memory without possessing explicit variability in criterion suggests that such variability influences performance in recognition nontrivially.

Evidence From Psychophysical Tasks

Thurstonian-type models with criterial variability have been more widely considered in psychophysics and psychoacoustics, where they have generally met with considerable success. Nosofsky (1983) found that range effects in auditory discrimination were due to increasing representational and criterial variance with wider ranges. Bonnel and Miller (1994) found evidence of considerable criterial variance in a same–different line-length judgment task in which attention to two stimuli was manipulated by instruction. They concluded that criterial variability was greater than representational variability in their task (see Bonnel & Miller, 1994, Experiment 2) and that focused attention served to decrease that variance.

Comparisons of Forced-Choice and Yes–No Procedures

One of the outstanding early successes of TSD is the proof by Green (1964; Green & Moses, 1966) that the area under the isosensitivity function as estimated by the rating-scale procedure should be equal to the proportion of correct responses in a two-alternative forced-choice task. This result generalizes across any plausible assumption about the shape of evidence distributions, as long as they are continuous, and is thus not limited by the assumption of normality typically imposed on TSD. Empirical verification of this claim would strongly support the assumptions underlying TSD, including that of a nonvariable criterion, but the extant work on this topic is quite mixed.

In perceptual tasks, this relationship appears to be approximately correct under some conditions (Emmerich, 1968; Green & Moses, 1966; Schulman & Mitchell, 1966; Shipley, 1965; Whitmore, Williams, & Ermey, 1968) but is not as strong or as consistent as one might expect (Lapsley Miller, Scurfield, Drga, Galvin, & Whitmore, 2002). Even within a generalization of Green's principle to a wide range of other decision axes and decision variables (Lapsley Miller et al., 2002), considerable observer inconsistency was noted. Such inconsistency is the province of our exploration here. In fact, a relaxation of the assumption of nonvariable criteria permits conditions in which this relationship can be violated. Wickelgren (1968) noted that it was "quite amazing" (p. 115) that the relationship appeared to hold even approximately.

The empirical evidence regarding the correspondence between forced-choice and yes–no recognition also suggests an inadequacy in the basic model. Green and Moses (1966) reported one experiment that conformed well to the prediction (Experiment 2) and one that violated it somewhat (Experiment 1). Most recent studies have made this comparison under the equal-variance assumption reviewed earlier as inadequate for recognition memory (Deffenbacher, Leu, & Brown, 1981; Khoe, Kroll, Yonelinas, Dobbins, & Knight, 2000; Yonelinas, Hockley, & Murdock, 1992), but experiments that have relaxed this assumption have yielded mixed results: Some have concluded that TSD-predicted correspondences are adequate (Smith & Duncan, 2004), and others have concluded in favor of other models (Kroll, Yonelinas, Dobbins, & Frederick, 2002). However, Smith and Duncan (2004) used rating scales for both forced-choice and yes–no recognition, making it impossible to establish whether their correspondences were good because ratings imposed no decision noise or because the criterion variance imposed by ratings was more or less equivalent on the two tasks. In addition, patients with amnesia, who might be expected to have a great difficulty with the maintenance of criteria, have been shown to perform relatively more poorly on yes–no than forced-choice recognition (Freed, Corkin, & Cohen, 1987; see also Aggleton & Shaw, 1996), although this result has not been replicated (Khoe et al., 2000; Reed, Hamann, Stefanacci, & Squire, 1997). The inconsistency in this literature may reflect the fact that criterion noise accrues throughout an experiment: Bayley, Wixted, Hopkins, and Squire (2008) recently showed that, whereas patients with amnesia do not show any disproportionate impairment on yes–no recognition on early testing trials, their performance on later trials does indeed drop relative to control subjects.

Although we do not pursue the comparison of forced-choice and yes–no responding further in our search for evidence of criterial variability, it is noteworthy that the evidence in support of the fundamental relationship between the two tasks reported by Green (1964) has not been abundant and that the introduction of criterial variability allows conditions under which that relationship is violated.

Recognition Memory and the Detection Formulation With Criterial Noise

This section outlines the mathematical formulation of the decision task and the basic postulates of TSD and extends that formulation by explicitly modeling criterion placement as a random variable with nonzero variability. To start, let us consider a subject's perspective on the task. Recognition requires the subject to discriminate between previously studied and unstudied stimuli. The traditional formulation of recognition presumes that test stimuli yield mnemonic evidence for studied status and that prior study affords discriminability between studied and unstudied stimuli by increasing the average amount of evidence provided by studied stimuli and likely increasing variance as well. However, inherent variability within both unstudied and studied groups of stimuli yields overlapping distributions of evidence. This theoretical formulation is depicted in Figure 1A, in which normal probability distributions represent the evidence values (e) that previously

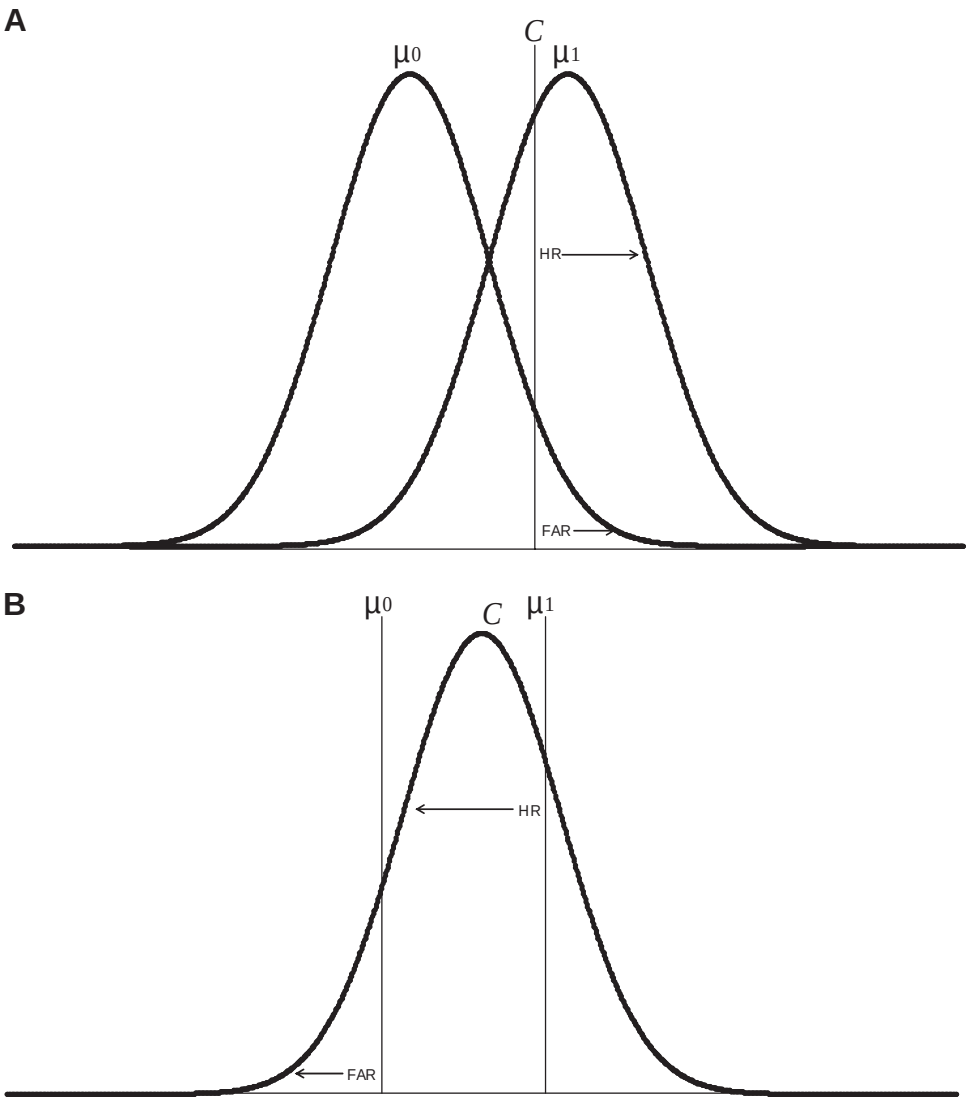

Figure 1. A: Traditional theory of signal detection representation of the recognition problem, including variable evidence distributions and a scalar criterion. B: An alternative formulation with scalar evidence values and a variable criterion. Both depictions lead to equivalent performance. *C* = decision criterion; FAR = false-alarm rate; HR = hit rate.

unstudied (S_0) and previously studied (S_1) stimuli yield at test.[4] If these distributions are nonzero over the full range of the evidence variable, then there is no amount of evidence that is unequivocally indicative of a particular underlying distribution (studied or unstudied). Equivalently, the likelihood ratio at criterion is $-\infty < \beta < \infty$. The response is made by imposing a decision criterion (c), such that

$$\begin{aligned} \text{Say ``Unstudied''} \quad & \text{if } (e - c) < 0. \\ \text{Say ``Studied''} \quad & \text{if } (e - c) \geq 0. \end{aligned}$$

The indicated areas in Figure 1 corresponding to hit and false-alarm rates (HR and FAR, respectively) illustrate how the variability of the representational distributions directly implies a particular level of performance.

Consider, as a hypothetical alternative case, a system without variable stimulus encoding. In such a system, signal and noise are represented by nonvariable (and consequently nonoverlapping) distributions of evidence, and the task seems trivial. Yet there is, in fact, some burden on the decision maker in this situation. First, the criterion must be placed judiciously—were it to fall anywhere

[4] Signal (studied status) and noise (unstudied status) distributions are referred to by the subscripts 1 and 0, respectively, throughout. This notation ensures more transparent generality to situations involving more than two distributions and can be thought of either as a dummy variable or as representing the number of presentations of the stimulus during the study phase.

outside the two evidence points, performance would be at chance levels. Thus, as reviewed previously, criterion placement must be a dynamic and feedback-driven process that takes into account aspects of the evidence distributions and the costs of different types of errors. Here, we explicitly consider the possibility that there is an inherent noisiness to criterion placement in addition to such systematic effects.

Figure 1B illustrates this alternative scenario, in which the decision criterion is a normal random variable with variance greater than 0 and *e* is a binary variable. Variability in performance in this scenario derives from variability in criterion placement from trial to trial but yields—in the case of this example—the same performance as in Figure 1A (shown by the areas corresponding to HR and FAR). This model fails, of course, to conform with our intuitions, and we show presently that it is untenable. However, the demonstration that criterial variability can yield outcomes identical with those of evidence variability is illustrative of the predicament we find ourselves in, namely, how to empirically distinguish between these two components of variability. The next section of this article outlines the problem explicitly.

Distribution of the Decision Variable and the Isosensitivity Function

Let μ_x and σ_x indicate the mean and standard deviation of distribution *x,* and let the subscripts *e* and *c* refer to evidence and criterion, respectively. If both evidence and criterial variability are assumed to be normally distributed (*N*) and independent of one another, as generally assumed by Thurstone (1927) and descendant models (Kornbrot, 1980; Peterson et al., 1954; Tanner & Swets, 1954), the decision variable is distributed as

$$(e - c) \sim N\left(\mu_e - \mu_c, \sqrt{\sigma_e^{\,2} + \sigma_C^2}\right). \tag{1a}$$

Because the variances of the component distributions sum to form the variability of the decision variable, it is not possible to discriminate between evidence and criterial variability on a purely theoretical basis (see also Wickelgren & Norman, 1966). This constraint does not preclude an empirical resolution, however. In addition, reworking the Thurstone model such that criteria cannot violate order constraints yields a model in which theoretical discrimination between criterion and evidence noise may be possible (Rosner & Kochanski, 2008).

Performance in a recognition task can be related to the decision variable by defining areas over the appropriate evidence function and, as is typically done in TSD, assigning the unstudied (e_0) distribution a mean of 0 and unit variance:

$$\mathrm{FAR} = p(\text{respond } S|e_0) = \int_{\mu_c}^{\infty} N\left(0,\ \sqrt{1 + \sigma_C^2}\right),$$

$$\mathrm{HR} = p(\text{respond } S|e_1) = \int_{\mu_c}^{\infty} N\left(\mu_1,\ \sqrt{\sigma_1^2 + \sigma_C^2}\right), \tag{1b}$$

in which *respond S* indicates a signal response, or a "yes" in a typical recognition task. These values are easiest to work with in normal-deviate coordinates:

$$z\mathrm{FAR} = \frac{-\mu_c}{\sqrt{1 + \sigma_C^2}},$$

$$z\mathrm{HR} = \frac{\mu_1 - \mu_c}{\sqrt{\sigma_1^2 + \sigma_C^2}}. \tag{1c}$$

Substitution and rearrangement yield the general model for the isosensitivity function with both representational and criterial variability (for related derivations, see McNicol, 1972; Wickelgren, 1968):

$$z\mathrm{HR} = \frac{\mu_1}{\sqrt{\sigma_1^2 + \sigma_C^2}} + z\mathrm{FAR}\frac{\sqrt{1 + \sigma_C^2}}{\sqrt{\sigma_1^2 + \sigma_C^2}}. \tag{2}$$

Note that, by this formulation, the slope of the function is not simply the reciprocal of the signal standard deviation, as it is in unequal-variance TSD. Increasing evidence variance will indeed decrease the slope of the function. However, the variances of the evidence and the criterion distribution also have an interactive effect: When the signal variance is greater than 1, increasing criterion variance will increase the slope. When it is less than 1, increasing criterion variance will decrease the slope. Equivalently, criterial variance reduces the effect of stimulus variance and pushes the slope toward 1.

Figure 2 depicts how isosensitivity functions vary as a function of criterial variance and confirms the claim of previous theorists (Treisman & Faulkner, 1984; Wickelgren, 1968) and implication of Equation 2 that criterial variability generally decreases the area under the isosensitivity function. The slight convexity at the margins of the function that results from unequal variances is an exception to that generality (see also Thomas and Myers, 1972). The left panels depict increasing criterial variance for signal variance less than 1, the right panels for signal variance greater than 1. The middle panels show that, when signal variance is equal to noise variance, criterial variance decreases the area under the curve but the slope does not change. It is worth noting that the prominent attenuating effect of criterial variance on the area under the function is generalizable across a number of plausible alternative distributions (including the logistic and gamma distributions; Thomas & Myers, 1972).

When criterial variability is zero, Equation 2 reduces to the familiar form of the unequal-variance model of TSD:

$$z\mathrm{HR} = \frac{\mu_1}{\sigma_1} + z\mathrm{FAR}\frac{1}{\sigma_1},$$

in which the slope of the function is the reciprocal of the signal variance and the y-intercept is μ_1/σ_1. When the distributions are assumed to have equal variance, as shown in Figure 1A, the slope of this line is 1.

When stimulus variability is zero and criterion variability is nonzero, as in Figure 1B, the isosensitivity function is

$$z\mathrm{HR} = \frac{\mu_1}{\sigma_C} + z\mathrm{FAR},$$

and when stimulus variability is nonzero but equal for the two distributions, the *z*ROC is

$$z\mathrm{HR} = \frac{\mu_1}{\sqrt{1 + \sigma_C^2}} + z\mathrm{FAR}.$$

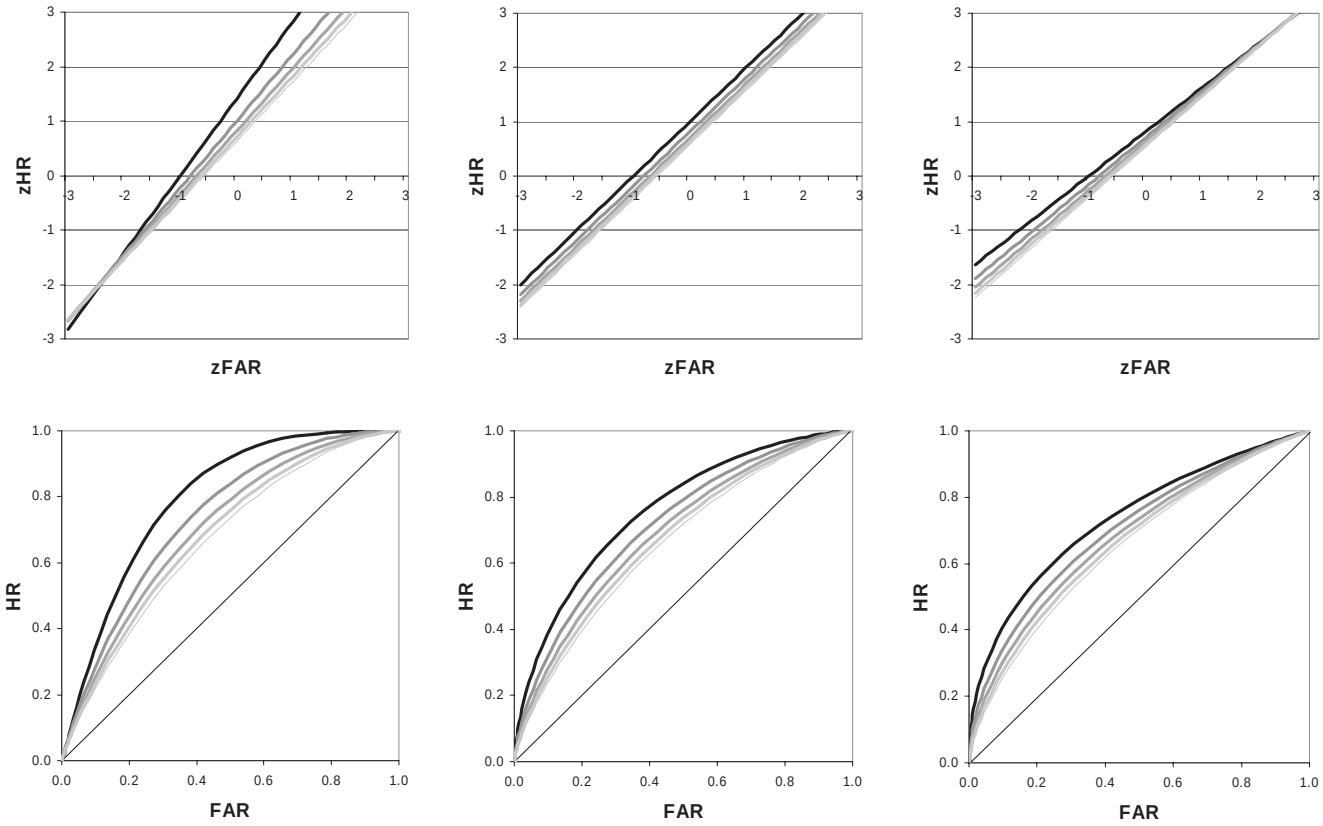

Figure 2. Isosensitivity functions in probability coordinates (bottom row) and normal-deviate coordinates (top row) for increasing levels of criterial noise (indicated by increasingly light contours). Left panels illustrate the case when the variability of the signal (old-item) distribution is less than that of the noise distribution (which has unit variance), middle panels when they are equal in variance, and right panels the (typical) case when the signal distribution is more variable than the noise distribution. FAR = false-alarm rate; HR = hit rate.

In both cases, the function has a slope of 1 and is thus identical to the case in which representational variability is nonzero but does not vary with stimulus type; thus, there is no principled way of using the isosensitivity function to distinguish between the two hypothetical cases shown in Figure 1, in which either evidence but not criterial variability or criterial but not evidence variability is present. Thankfully, given the actual form of the empirical isosensitivity function—which typically reveals a nonunit slope—we can use the experimental technique presented later in this article to disentangle these two bases.

Measures of Accuracy With Criterion Variability

Because empirical isosensitivity functions exhibit nonunit slope, we need to consider measures of accuracy that generalize to the case when evidence distributions are not of equal variance. This section provides the rationale and derivations for ND-TSD generalizations of two commonly used measures, d_a and d_e.

There are three basic ways of characterizing accuracy (or, variously, discriminability or sensitivity) in the detection task. First, accuracy is related to the degree to which the evidence distributions overlap and is thus a function of the distance between them, as well as their variances. Second, accuracy is a function of the distance of the isosensitivity line from an arbitrary point on the line that represents complete overlap of the distributions (and thus chance levels of accuracy on the task). Finally, accuracy can be thought of as the amount of area below an isosensitivity line—an amount that increases to 1 when performance is perfect and drops to 0.5 when performance is at chance. Each of these perspectives has interpretive value: The distribution-overlap conceptualization is easiest to relate to the types of figures associated with TSD (e.g., Figure 1A); distance-based measures emphasize the desirable psychometric qualities of the statistic (e.g., that they are on a ratio scale; Matzen & Benjamin, in press). Area-based measures bear a direct and transparent relation with forced-choice tasks. All measures can be intuitively related to the geometry of the isosensitivity space.

To derive measures of accuracy, we deal with the distances from the isosensitivity line, as defined by Equation 2.[5] Naturally, there are an infinite number of distances from a point to a line, so it is necessary to additionally restrict our definition. Here, we do so by using the shortest possible distance from the origin to the line,

[5] For an intuitive and thorough review of the geometry underlying detection parameters, see Wickens (2002).

which yields a simple linear transformation of d_a (Schulman & Mitchell, 1966). In Appendix A, we provide an analogous derivation for d_e, which is the distance from the origin to the point on the isosensitivity line that intersects with a line perpendicular to the isosensitivity line. These values also correspond to distances on the evidence axis scaled by the variance of the underlying evidence distributions: d_e corresponds to the distance between the distributions, scaled by the arithmetic average of the standard deviations, and d_a corresponds to distance in terms of the root-mean-square average of the standard deviations (Macmillan & Creelman, 2005). For the remainder of this article, we use d_a, as it is quite commonly used in the literature (e.g., Banks, 2000; Matzen & Benjamin, in press), is easily related to area-based measures of accuracy, and provides a relatively straightforward analytic form.

The generalized version of d_a can be derived by solving for the point at which the isosensitivity function must intersect with a line of slope $(-1/m)$:

$$z\mathrm{HR} = -z\mathrm{FAR}\left(\frac{\sqrt{\sigma_1^2 + \sigma_c^2}}{\sqrt{1 + \sigma_c^2}}\right).$$

The intersection point of Equation 2 and this equation is

$$\left(\frac{-\mu_1\sqrt{1+\sigma_c^2}}{\sigma_1^2 + 2\sigma_c^2 + 1}, \frac{\mu_1\sqrt{\sigma_1^2 + \sigma_c^2}}{\sigma_1^2 + 2\sigma_c^2 + 1}\right),$$

which yields a distance of

$$\text{noisy } d_a^* = \frac{\mu_1\sqrt{\sigma_1^2 + 2\sigma_c^2 + 1}}{\sigma_1^2 + 2\sigma_c^2 + 1}$$

from the origin. This value is scaled by $\sqrt{2}$ to determine the length of the hypotenuse on a triangle with sides of length noisy d_a^* (Simpson & Fitter, 1973):

$$\text{noisy } d_a = \frac{\sqrt{2}\mu_1\sqrt{\sigma_1^2 + 2\sigma_c^2 + 1}}{\sigma_1^2 + 2\sigma_c^2 + 1}. \quad (3)$$

The area measure A_Z also bears a simple relationship with d_a^*:

$$\text{noisy } A_Z = \Phi\,[\text{noisy } d_a^*].$$

Empirical Estimation of Sources of Variability

Because both criterial variability and evidence variability affect the slope of the isosensitivity function, it is difficult to isolate the contributions of each to performance. To do so, we must find conditions over which we can make a plausible case for criterial and evidential variance being independently and differentially related to a particular experimental manipulation. We start by taking a closer look at this question.

Units of Variability for Stimulus and Criterial Noise

Over what experimental factor is evidence presumed to vary? Individual study items probably vary in preexperimental familiarity and also in the effect of a study experience. In addition, the waxing and waning of attention over the course of an experiment increase the item-related variability (see also DeCarlo, 2002).

Do these same factors influence criterial variability? By the arguments presented here, criterial variability related to item characteristics is mostly systematic in nature (see, e.g., Benjamin, 2003) and is thus independent of the variability modeled by Equation 1. We have specifically concentrated on nonsystematic variability and have argued that it is likely a consequence of the cognitive burden of criterion maintenance. Thus, the portion of criterial variability with which we concern ourselves with is trial-to-trial variability on the test. What is needed is a paradigm in which item variability can be dissociated from trial variability.

Ensemble Recognition

In the experiment reported here, we used a variant of a clever paradigm devised by Nosofsky (1983) to investigate range effects in the absolute identification of auditory signals. In our experiment, subjects made recognition judgments for ensembles of items that varied in size. Thus, each test stimulus included a variable number of words (one, two, or four), all of which were old or all of which were new. The subjects' task was to evaluate the ensemble of items and provide an "old" or "new" judgment on the group.

The size manipulation is presumed to affect stimulus noise (because each ensemble is composed of heterogeneous stimuli and is thus subject to item-related variance) but not criterial noise (because the items are evaluated within a single trial, as a group). Naturally, this assumption might be incorrect: Subjects might, in fact, evaluate each item in an ensemble independently and with heterogeneous criteria. We examine the data closely for evidence of a violation of the assumption of criterial invariance within ensembles.

Information Integration

To use the data from ensemble recognition to separately evaluate criterial and stimulus variance, we must have a linking model of information integration within an ensemble—that is, a model of how information from multiple stimuli is evaluated jointly for the recognition decision. We consider two general models. The *independent variability model* proposes that the variance of the strength, but not the criterial distribution, is affected by ensemble size, as outlined in the previous section. Four submodels are considered. The first two assume that evidence is averaged across the stimuli within an ensemble and differ only in whether criterial variability is permitted to be nonzero (ND-TSD) or not (TSD). The latter two assume that evidence is summed across the stimuli within an ensemble and, as before, differ in whether criterion variability is allowed to be nonzero. These models will be compared with the OR model, which proposes that subjects respond positively to an ensemble if any member within that set yields evidence greater than a criterion. This latter model embodies a failure of the assumption that the stimuli are evaluated as a group, and its success would imply that our technique for separating criterial and stimulus noise is invalid. Thus, five models of information integration are considered.

Criterion Placement

For each ensemble size, five criteria had to be estimated to generate performance on a 6-point rating curve. For all models except the two summation models, a version of the model was fit in which criteria were free to vary across ensemble size (yielding

15 free parameters and henceforth referred to as *without restriction*), and another version was fit in which the criteria were constrained (*with restriction*) to be the same across ensemble sizes (yielding only five free parameters). Because the scale of the mean evidence values varies with ensemble size for the summation models, only one version was fit, in which there were 15 free parameters (i.e., they were free to vary across ensemble size).

Model Flexibility

One important concern in comparing models, especially nonnested models like the OR model, is that a model may benefit from undue flexibility. That is, a model may account for a data pattern more accurately not because it is a more accurate description of the underlying generating mechanisms but rather because its mathematical form affords it greater flexibility (Myung & Pitt, 2002). It may thus appear superior to another model by virtue of accounting for nonsystematic aspects of the data. There are several approaches we have taken to reduce concerns that ND-TSD may benefit from greater flexibility than its competitors.

First, we have adopted the traditional approach of using an index of model fit that is appropriate for nonnested models and penalizes models according to the number of their free parameters (the Akaike information criterion [AIC]; Akaike, 1973). Second, we use a correction on the generated statistic that is appropriate for the sample sizes in use here (AIC_c; Burnham & Anderson, 2004). Third, we additionally report the Akaike weight metric, which, unlike the AIC or AIC_c, has a straightforward interpretation as the probability that a given model is the best among a set of candidate models. Fourth, in addition to reporting both AIC_c values and Akaike weights, we also report the number of subjects best fit by each model, ensuring that no model is either excessively penalized for failing to account for only a small number of subjects (but dramatically so) or bolstered by accounting for only a small subset of subjects considerably more effectively than the other models.

Finally, we report in Appendix C the results of a large series of Monte Carlo simulations evaluating the degree to which ND-TSD has an advantage over TSD in terms of accounting for failures of assumptions common to the two models. We consider cases in which the evidence distributions are of a different form than assumed by TSD and cases in which the decision rule is different from what we propose. To summarize the results from that exercise here, ND-TSD never accrues a higher AIC score or Akaike weight than TSD unless the generating distribution is ND-TSD itself. These results indicate that a superior fit of ND-TSD to empirical data is unlikely to reflect undue model flexibility when compared with TSD.

Experiment: Word Ensemble Recognition

In this experiment, we evaluated the effects of manipulating study time on recognition of word ensembles of varying sizes. By combining ND-TSD and TSD with a few simple models of information integration, we were able to separately estimate the influence of criterial and evidence variability on recognition across those two study conditions. This experiment pit the models outlined in the previous section against one another.

Method

Subjects. Nineteen undergraduate students from the University of Illinois at Urbana–Champaign participated to partially fulfill course requirements for an introductory course in psychology.

Design. Word set size (one, two, or four words in each set) was manipulated within subjects. Each subject participated in a single study phase and a single test phase. Subjects made their recognition responses on a 6-point confidence-rating scale, and the raw frequencies of each response type were fit to the models to evaluate performance.

Materials. All words were obtained from the English Lexicon Project (Balota et al., 2002). We drew 909 words with a mean word length of 5.6 (range: 4–8 letters) and mean log HAL frequency of 10.96 (range: 5.5–14.5). A random subset of 420 words was selected for the test list, which consisted of 60 single-item sets, 60 double-item sets, and 60 four-item sets. A random half of the items from each ensemble size set was assigned to the study list. All study items were presented singly, while test items were presented in sets of one, two, or four items. Words presented in a single ensemble were either all previously studied or all unstudied. This resulted in 210 study item presentations and 180 test item presentations (90 old and 90 new). Again, every test presentation included all old or all new items; there were no trials on which old and new items were mixed in an ensemble.

Procedure. Subjects were tested individually in a small, well-lit room. Stimuli were presented, and subject responses were recorded, on PC-style computers programmed using the Psychophysical Toolbox for MATLAB (Brainard, 1997; Pelli, 1997). Prior to the study phase, subjects read instructions on the computer screen informing them that they were to be presented with a long series of words that they were to try and remember as well as they could. They began the study phase by pressing the space bar. During the study phase, words were presented for 1.5 s. There was a 333-ms interstimulus interval (ISI) between presentations. At the conclusion of the study phase, subjects were given instructions for the test phase. Subjects were informed that test items would be presented in sets of one, two, or four words and that they were to determine if the word or words that they were presented had been previously studied or not. They began the test phase by pressing the space bar. There was no time limit on the test.

Results

Table 1 shows the frequencies by test condition summed across subjects. Discriminability (d_a) was estimated separately for each ensemble size and study time condition by maximum-likelihood estimation (Ogilvie & Creelman, 1968), and subject means are displayed in Table 1. All model fitting reported below was done on the data from individual subjects because of well-known problems with fitting group data (see, e.g., Estes & Maddox, 2005) and particular problems with recognition data (Heathcote, 2003) None of the subjects or individual trials were omitted from analysis. Details of the fitting procedure are outlined in Appendix D.

The subject-level response frequencies were used to evaluate the models introduced earlier. Of particular interest is the independent variability model that we use to derive separate estimates of criterial and evidence variability. That model's performance is

Table 1
Summed Rating Frequencies Over Ensemble Size and Old–New Test Items

Condition	1	2	3	4	5	6	μ_1	σ_1	m	d_a
Size = 1										
New	105	109	124	105	67	60	0.75	1.28	0.78	0.64
Old	62	65	64	108	103	168				
Size = 2										
New	107	139	115	99	70	40	1.25	1.52	0.66	0.88
Old	47	67	81	80	82	213				
Size = 4										
New	134	146	108	76	57	49	1.50	1.60	0.62	1.10
Old	53	54	61	56	104	242				

Note. Standard unequal-variance estimates of model parameters are shown on the right. μ_1 = mean of the signal distribution; σ_1 = standard deviation of the signal distribution; m = slope of the isosensitivity function; d_a = a standard estimate of memory sensitivity.

evaluated with respect to several other models. One is a submodel (zero criterial variance model) that is equivalent to the independent variability model but assumes no criterial variance. For both the model with criterion variability (ND-TSD) and without (TSD), two different decision rules (averaging vs. summation) were tested. Another model (the OR model) assumes that each stimulus within an ensemble is evaluated independently and that the decision is made on the basis of combining those independent decisions via an OR rule. Comparison of the independent variability model with the OR model was used to evaluate the claim that the stimulus is evaluated as an ensemble, rather than as n individual items. Comparison of the independent variability model with the nested zero criterial-variance model was used to test for the presence of criterial variability.

Independent variability models. The averaging version of this model is based on ND-TSD and the well-known relationship between the sampling distribution of the mean and sample size, as articulated by the central limit theorem. Other applications of a similar rule in psychophysical tasks (e.g., Swets & Birdsall, 1967; Swets, Shipley, McKey, & Green, 1959) have confirmed this assumption of averaging stimuli or samples, but we evaluate it carefully here because of the novelty of applying that assumption to recognition memory.

If the probability distribution of stimulus strength has variability σ^2, then that probability distribution for the ensemble of n stimuli drawn from that distribution has variability σ^2/n. This model assumes that the distribution of strength values is affected by n but that criterial variability is not. Thus, the isosensitivity function of the criterion variance ensemble recognition model is

$$z\mathrm{HR} = \frac{\mu_1}{\sqrt{\frac{\sigma_1^2}{n} + \sigma_C^2}} + z\mathrm{FAR}\frac{\sqrt{\frac{1}{n} + \sigma_C^2}}{\sqrt{\frac{\sigma_1^2}{n} + \sigma_C^2}}. \quad (4)$$

Because we fit the frequencies directly rather than using the derived estimates of distance (unlike previous work: Nosofsky, 1983), there was no need to fix any parameters (such as the distance between the distributions) a priori. The hypothesized effect of the ensemble size manipulation is shown in Figure 3, in which the variance of the stimulus distributions decreases with increasing size. For clarity, the criterion distribution is not shown.

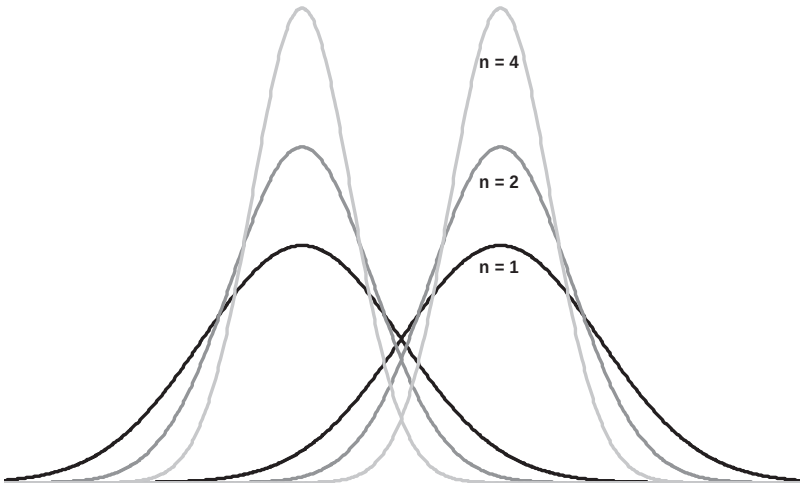

Figure 3. Predictions of the variability models of information integration for the relationship between ensemble size (n) and the shapes of the evidence distributions.

The fit of this model was compared with a simpler model that assumes no criterial variability:

$$z\mathrm{HR} = \frac{\mu_1}{\sqrt{\frac{\sigma_1^2}{n}}} + z\mathrm{FAR}\frac{1}{\sigma_1}. \quad (5)$$

Another possibility is that evidence is summed, rather than averaged, within an ensemble. In this case, the size of the ensemble scales both the signal mean and the stimulus variances, and the isosensitivity functions assumes the form

$$z\mathrm{HR} = \frac{n\mu_1}{\sqrt{n\sigma_1^2 + \sigma_C^2}} + z\mathrm{FAR}\frac{\sqrt{n + \sigma_C^2}}{\sqrt{n\sigma_1^2 + \sigma_C^2}} \quad (6)$$

when criterion variance is nonzero and

$$z\mathrm{HR} = \frac{\sqrt{n}\mu_1}{\sigma_1} + z\mathrm{FAR}\frac{1}{\sigma_1} \quad (7)$$

when criterion variance is zero. Note that Equations 7 and 5 are equivalent, demonstrating that the summation rule is equivalent to the averaging rule when criterion variability is zero.

We must also consider the possibility that our assumption of criterial invariance within an ensemble is wrong. If criterial variance is affected by ensemble size in the same purely statistical manner as is stimulus variance, then both stimulus and criterion variance terms are affected by *n*. Under these conditions, the model is

$$z\mathrm{HR} = \frac{\sqrt{n}\mu_1}{\sqrt{\sigma_1^2 + \sigma_C^2}} + z\mathrm{FAR}\frac{\sqrt{\frac{1 + \sigma_C^2}{n}}}{\sqrt{\frac{\sigma_1^2 + \sigma_C^2}{n}}}. \quad (8)$$

Two aspects of this model are important. First, it can be seen that it is impossible to separately estimate the two sources of variability because they can be combined into a single superparameter. Second, as shown in Appendix B, this model reduces to the same form as Equation 5 and thus can fit the data no better than the zero criterial variability model. Consequently, if the zero criterial variability model is outperformed by the independent variability model, then we have supported the assumption that criterial variability is invariant across an ensemble.

OR model. The OR model assumes that each stimulus within an ensemble is evaluated independently and that subjects respond positively to a set if any one of those stimuli surpasses a criterion value (e.g., Macmillan & Creelman, 2005; Wickens, 2002). This is an important baseline against which to evaluate the information integration models because the interpretation of those models hinges critically on the assumption that the ensemble manipulation alters representational variability in predictable ways embodied by Equations 4–8. The OR model embodies a failure of this assumption: If subjects do not average or sum evidence across the stimuli in an ensemble but rather evaluate each stimulus independently, then this multidimensional extension of the standard TSD model will provide a superior fit to the data.

The situation is simplified because the stimuli within an ensemble (and, in fact, across the entire study set) can be thought of as multiple instances of a common random variable. The advantage of this situation is apparent in Figure 4, which depicts the two-dimensional TSD representation of the OR model applied to two stimuli. Here, the strength distributions are shown jointly as density contours; the projection of the marginal distributions onto either axis represents the standard TSD case. Because the stimuli are represented by a common random variable, those projections are equivalent.

According to the standard TSD view, a subject provides a rating of *r* to a stimulus if and only if the evidence value yielded by that stimulus exceeds the criterion associated with that rating, C_r. Thus, the probability of at least one of *n* independent and identically distributed instances of that random variable exceeding that criterion is

$$p(e_n > C_r) = 1 - [p(e < C_r)]^n = 1 - [1 - p(e \geq C_r)]^n. \quad (9)$$

The unshaded portion of the figure corresponds to the bracketed term in Equation 9. The region of endorsement for a subject is the shaded area (above either criterion, and extends leftward and downward to $-\infty$).

Model fitting. Details of the model-fitting procedure are provided in Appendix D.

Model results. The performance of the models is shown in Table 2, which indicates AIC_C, Akaike weights, and number of individual subjects best fit by each model. It is clear that the superior fit was provided by ND-TSD with the restriction of equivalent criteria across ensemble conditions and with the averaging rather than the summation process. That model provided the best fit (lowest AIC_c score) for more than 80% of the individual subjects and had (on average across subjects) a greater than 80% chance of being the best model in the set tested. This result is consistent with the presence of criterial noise and additionally with the suggestion that subjects have a very difficult time adjusting criteria across trials (e.g., Ratcliff & McKoon, 2000).

A depiction of the fit of the winning model is shown in the top panel of Figure 5, in which it can be seen that ND-TSD provides quite a different conceptualization of the recognition process than

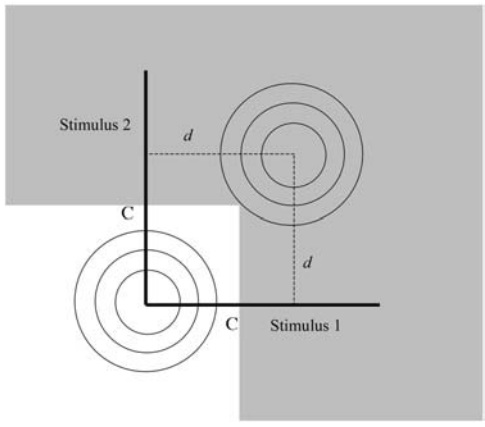

Figure 4. The multidimensional formulation of the OR model for information integration. Distributions are shown from above. Given a criterion *C* and discriminability *d* for a single stimulus, the shaded area represents predicted performance on the joint stimulus.

Table 2
Corrected Akaike Information Criterion Values (AIC_c), Akaike Weights, and Number of Subjects Best Fit by Each Model

Model and parameter	Without restriction	With restriction
ND-TSD averaging model		
Mean AIC_c	154	**116**
Mean Akaike weight	0.00	**0.82**
Number of subjects	0	**16**
TSD averaging model		
Mean AIC_c	146	128
Mean Akaike weight	0.00	0.13
Number of subjects	0	3
ND-TSD summation model		
Mean AIC_c	152	
Mean Akaike weight	0.00	
Number of subjects	0	
TSD summation model		
Mean AIC_c	150	
Mean Akaike weight	0.05	
Number of subjects	1	
OR model		
Mean AIC_c	146	159
Mean Akaike weight	0.00	0.00
Number of subjects	0	0

Note. With restriction and without restriction columns refer to models in which criteria were allowed to vary across ensemble size (without restriction) or were not (with restriction). Parameters from the winning model are depicted in boldface. ND-TSD = noisy decision theory of signal detection; OR = model with "OR" decision rule; TSD = theory of signal detection.

does standard TSD (shown in the bottom panel). In addition to criterial variance, the variance of the studied population of items is estimated to be much greater relative to the unstudied population. This suggests that the act of studying words may confer quite substantial variability and that criterial variance acts to mask that variability. The implications of this are considered in the next major section.

Psychological Implications of Criterial Variance

When interpreted in the context of TSD, superior performance in one condition versus another or as exhibited by one subject over another is attributable either to a greater distance between the means of the two probability distributions or to lesser variability of the distributions. In ND-TSD, superior performance can additionally reflect lower levels of criterial variability. In this section, we outline several current and historical problems that may benefit from an explicit consideration of criterial variability. The first two issues we consider underlie current debates about the relationship between the slope of the isosensitivity function and theoretical models of recognition and of decision making. The third issue revisits the standoff between deterministic and probabilistic response models and demonstrates how decision noise can inform that debate. The fourth, fifth, and sixth issues address the effects of aging and the consequences of fatigue and consider the question of how subjects make introspective remember–know judgments in recognition tasks. These final points are all relevant to current theoretical and empirical debates in recognition memory.

Effects of Recognition Criterion Variability on the Isosensitivity Function

Understanding the psychological factors underlying the slope of the isosensitivity function have proven to be somewhat of a puzzle in psychology in general and in recognition memory in particular. Different tasks appear to yield different results: For example, recognition of odors yields functions with slopes ~1 (Rabin & Cain, 1984; Swets, 1986a), whereas recognition of words typically yields considerably shallower slopes (Ratcliff et al., 1992, 1994). That latter result is particularly important because it is inconsistent with a number of prominent models of recognition memory (Eich, 1982; Murdock, 1982; Pike, 1984). The form of the isosensitivity function has even been used to explore variants of recognition memory, including memory for associative relations (Kelley & Wixted, 2001; Rotello, Macmillan, & Van Tassel, 2000) and memory for source (M. R. Healy et al., 2005; Hilford, Glanzer, Kim, & DeCarlo, 2002).

One claim about the slope of the isosensitivity function in recognition memory is the *constancy-of-slopes generalization* and owes to the pioneering work of Ratcliff and his colleagues (1992, 1994), who found that slopes were not only consistently less than unity but also relatively invariant with manipulations of learning.

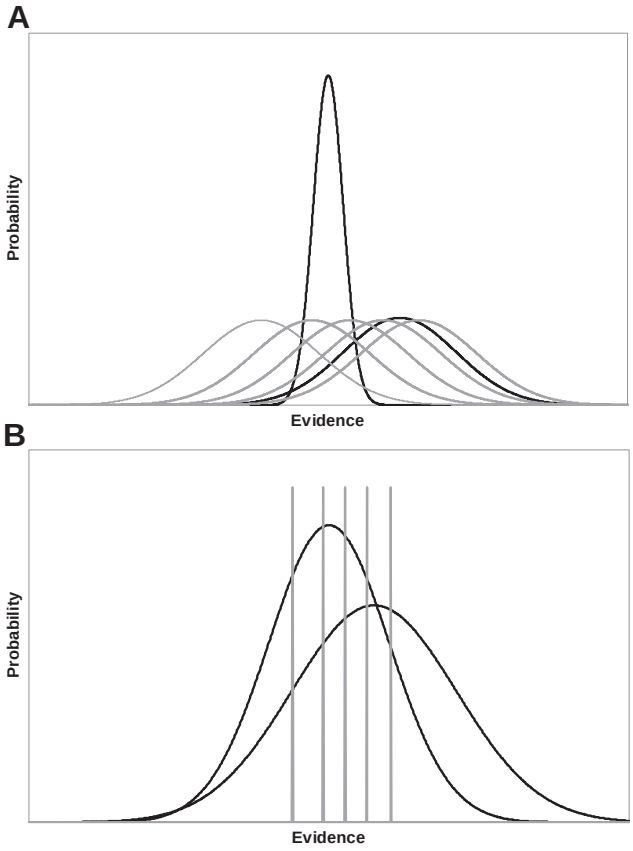

Figure 5. Depiction of the results from the winning noisy decision theory of signal detection model (Panel A) and traditional theory of signal detection (Panel B). Dark lines are evidence distributions, and lighter lines represent criteria.

Later work showed, however, that this may not be the case (Glanzer et al., 1999; Heathcote, 2003; Hirshman & Hostetter, 2000). In most cases, it appears as though variables that increase performance decrease the slope of the isosensitivity function (for a review, see Glanzer et al., 1999). This relation holds for manipulations of normative word frequency (Glanzer & Adams, 1990; Glanzer et al., 1999; Ratcliff et al., 1994), concreteness (Glanzer & Adams, 1990), list length (Elam, 1991, as reported in Glanzer et al., 1999; Gronlund & Elam, 1994; Ratcliff et al., 1994; Yonelinas, 1994), retention interval (Wais, Wixted, Hopkins, & Squire, 2006), and study time (Glanzer et al., 1999; Hirshman & Hostetter, 2000; Ratcliff et al., 1992, 1994).

These two findings—slopes of less than 1 and decreasing slopes with increasing performance—go very much hand in hand from a measurement perspective. Consider the limiting case, in which learning has been so weak and memory thus so poor, that discrimination between the old and new items on a recognition test is nil. The isosensitivity function must have a slope of 1 in both probability and normal-deviate coordinates in that case because any change in criterion changes the HR and FAR by the same amount. As that limiting case is approached, it is thus not surprising that slopes move toward 1. The larger question in play here is whether the decrease in performance that elicits that effect owes specifically to shifting evidence distributions or whether criterial variance might also play a role. We tackle this question below by carefully examining the circumstances in which a manipulation of learning affects the slope and the circumstances in which it does not.

The next problem we consider is why isosensitivity functions estimated from the rating task differ from those estimated by other means and whether such differences are substantive and revealing of fundamental problems with TSD. In doing so, we consider what role decision noise might play in promoting such differences and also whether reports of the demise of TSD (Balakrishnan, 1998a) may be premature.

Inconsistent effects of memory strength on slope. The first puzzle we consider concerns the conflicting reports on the effects of manipulations of learning on the slope of the isosensitivity function. Some studies have revealed that the slope does not change with manipulations of learning (Ratcliff et al., 1992, 1994), whereas others have supported the idea that the slope decreases with additional learning or memory strength. While some models of recognition memory predict changes in slope (Gillund & Shiffrin, 1984; Hintzman, 1986) with increasing memory strength, others either predict unit slope (Murdock, 1982) or invariant slope with memory strength. This puzzle is exacerbated by the lack of entrenched theoretical mechanisms that offer a reason why the effect should sometimes obtain and sometimes not.

To understand the way in which criterion noise might underlie this inconsistency, it is important to note the conditions under which changes in slope are robust and the conditions under which they are not. Glanzer et al. (1999) reviewed these data, and their results provide an important clue. Of the four variables for which a reasonable number of data were available ($\geq$5 independent conditions), list length and word frequency manipulations clearly demonstrated the effect of learning on slope: Shorter list lengths and lower word frequency led to higher accuracy and also exhibited a lower slope (in 94% of their comparisons). In contrast, greater study time and more repetitions led to higher accuracy but revealed the effect on slope less consistently (on only 68% of the comparisons).

To explore this discrepancy, we consider the criterion-setting strategies that subjects bring to bear in recognition and how different manipulations of memory might interact with those strategies. There are two details about the process of criterion setting and adjustment that are informative. First, the control processes that adjust criteria are informed by an ongoing assessment of the properties of the testing regimen. This may include information based on direct feedback (Dorfman & Biderman, 1971; Kac, 1962; Thomas, 1973, 1975) or derived from a limited memory store of recent experiences (Treisman, 1987; Treisman & Williams, 1984). In either case, criterion placement is likely to be a somewhat noisy endeavor until a steady state is reached, if it ever is. From the perspective of these models, it is not surprising that support has been found for the hypothesis that subjects set a criterion as a function of the range of experienced values (Parducci, 1984), even in recognition memory (Hirshman, 1995). These theories have at their core the idea that recognizers hone in on optimal criterion placement by assessing, explicitly or otherwise, the properties of quantiles of the underlying distributions. Because this process is subject to a considerable amount of irreducible noise—for example, from the particular order in which early test stimuli are received—decision variability is a natural consequence. To the degree that criterion variability is a function of the range of sampled evidence values (cf. Nosofsky, 1983), criterion noise will be greater when that range is larger.

The second relevant aspect of the criterion-setting process is that it takes advantage of the information conveyed by the individual test stimuli. A stimulus may reveal something about the degree of learning a prior exposure would have afforded it, and subjects appear to use this information in generating an appropriate criterion (J. Brown et al., 1977). Such a mechanism has been proposed as a basis for the mirror effect (Benjamin, 2003; Benjamin, Bjork, & Hirshman, 1998; Hirshman, 1995) and, according to such an interpretation, reveals the ability of subjects to adjust criteria on an item-by-item basis in response to idiosyncratic stimulus characteristics. It is noteworthy that within-list mirror effects are commonplace for stimulus variables, such as word frequency, meaningfulness, and word concreteness (Glanzer & Adams, 1985; 1990), but typically absent for experimental manipulations of memory strength, such as repetition (Higham et al., in press; Stretch & Wixted, 1998) or study time (Verde & Rotello, 2007). This difference has been taken to imply that recognizers are not generally willing or able to adjust criteria within a test list on the basis of an item's perceived strength class.

In fact, the few examples of within-list mirror effects arising in response to a manipulation of strength are all ones in which the manipulation provided for a relatively straightforward assignment to strength class, including variable study–test delay (Singer, Gagnon, & Richards, 2002; Singer & Wixted, 2006) and the use of stimuli that were associatively categorized (Benjamin, 2001; Starns, Hicks, & Marsh, 2006). Similar within-list manipulations of strength tied to color (Stretch & Wixted, 1998) or list half at test (Verde & Rotello, 2007) were unsuccessful, supporting the view that the relationship between the strength manipulation and the stimulus must be extremely transparent to support explicit differentiation by the subject.

What does this imply for the placement and maintenance of criteria across conditions that vary in discriminability? When the burden of assigning a test stimulus to a subclass falls on the recognizer, he or she will often forgo that decision. In that case, the recognizer will accumulate information on a single class of old items as he or she samples from the test stimuli. However, when the task relieves the subject of this burden, either by dividing up the discriminability classes between subjects or between test lists or by using stimuli that carry with them inherent evidence as to their appropriate class and likely discriminability, then the subject may treat as separate the estimation of range for the different classes.

The effects of these strategic differences can be seen in Figures 6 and 7. As shown in Figure 6, if increases in discriminability lead to increased stimulus variance and criterion noise is constant, the slope of the isosensitivity function should always decrease when conditions afford superior memory discrimination. This is shown in Boxes B and C. However, as criterion variance increases, the effect of stimulus variance becomes less pronounced (as can be seen by comparing the two boxes). Consider the effect of a manipulation of memory on the slope:

$$\text{effect} = \frac{\sqrt{1+\sigma_C^2}}{\sqrt{\sigma_1^2+\sigma_C^2}} - \frac{\sqrt{1+\sigma_C^2}}{\sqrt{\sigma_2^2+\sigma_C^2}},$$

in which the subscripts 1 and 2 denote the two levels of the manipulated variable, with level 2 being the condition with superior performance and greater stimulus variability. Under these conditions, it is easy to see that the value of this effect must be either 0 or positive. That is, if stimulus variance increases with discriminability, then the condition with greater discriminability must have a lower slope. The inconsistency in the literature must then come from the effect of those variables on criterion variability, which can attenuate the magnitude of the difference. When test stimuli are not successfully subclassified, then criterion variance reflects the full range of the old stimuli rather than the ranges of the individual classes.

Figure 7 illustrates the decision milieu that yields these differential effects. When the set of old items is heterogeneous with respect to discriminability but subjects do not discriminate between the strength classes, then the sampled range of criterion values reflects the full range of this mixture distribution, and the variability of the criterion will be great (shown in Figure 7B as the root-mean-square average of the criterion distributions in Figure 7A). When subjects do discriminate between the strength classes and sampled values from each class inform a unique criterion distribution (as shown in Figure 7A), then those two distributions will both be of lesser variability.

In both cases, criterion variance is constant across stimulus classes, and the net effect is a decrease in slope. This occurs because increasing stimulus variability is offset by a constant amount of criterion variance. However, criterion variance serves to effectively augment or retard the magnitude of the decrease. In Figure 7A, in which each item class specifies a unique criterion distribution, the lesser variability that accompanies the stimulus classes translates into a lesser amount of criterial variability. In this case, that lesser variability increases the degree to which stimulus variability yields an effect on slope.

By this explanation, memory-enhancing conditions that afford subclassification of the test stimuli with respect to discriminability should be more likely to yield an effect on slope than variables that are opaque with respect to discriminability. Now we are in a position to reconsider the empirically studied variables enumerated earlier. Variables that are manipulated between subjects or between lists require no subclassification within a test list and should thus provide for relatively easy assignment at test. Of the four variables mentioned earlier, only list length is always studied between lists (by definition). In addition, variables for which the discriminability class is inherent to the stimulus itself should also

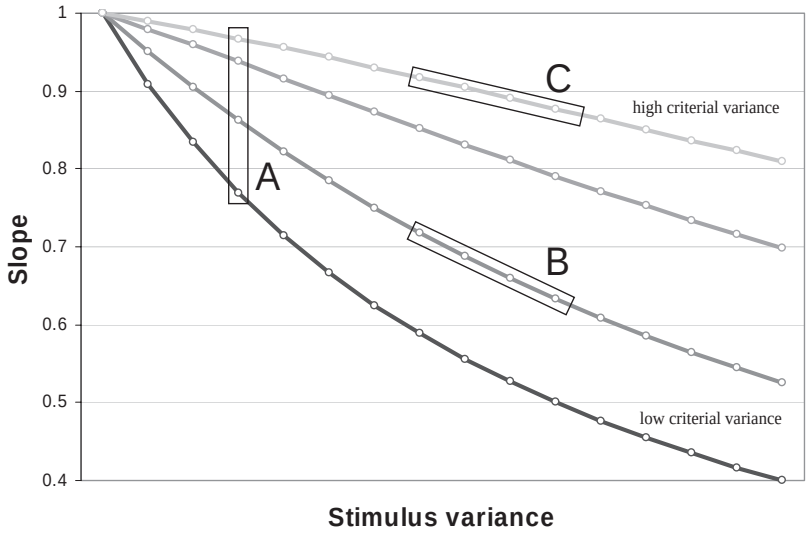

Figure 6. Slope of the isosensitivity curve as a function of stimulus (ranging from 1 to 2.5) and criterion variance (ranging from 0 to 3).

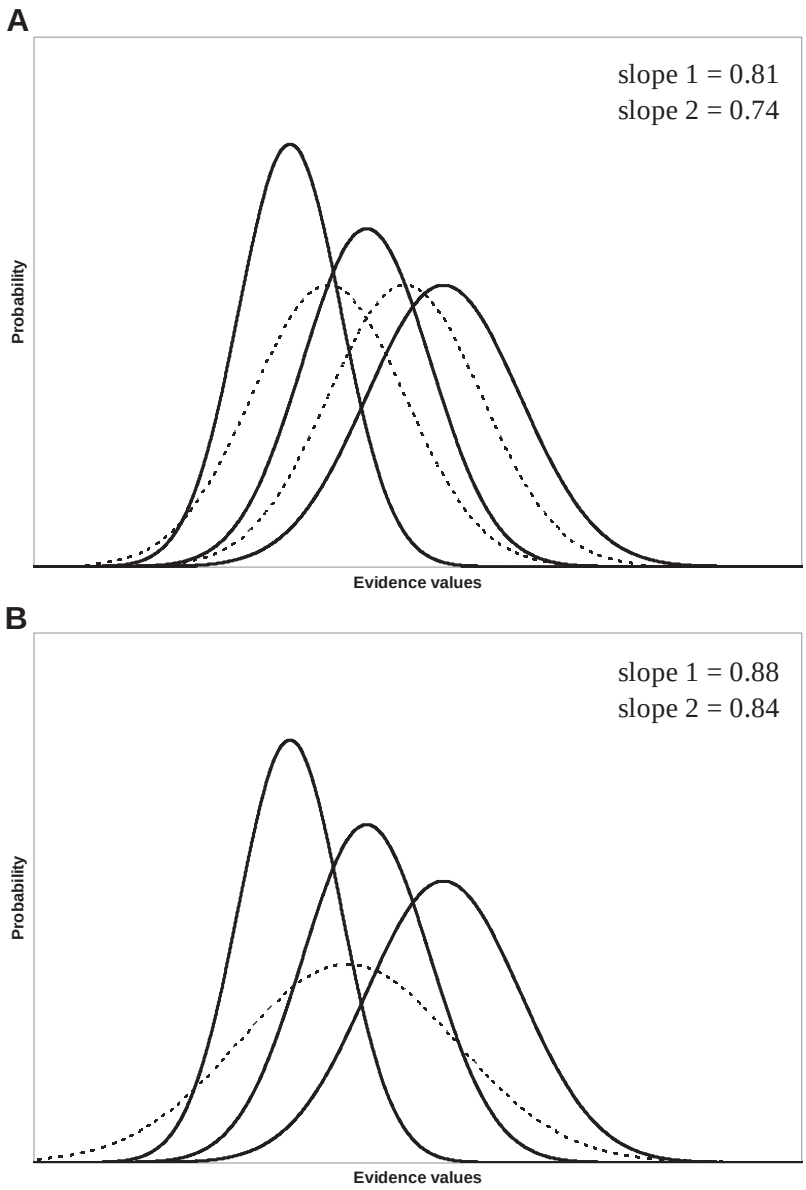

Figure 7. A demonstration of how a manipulation of learning can lead to a larger difference in slope between conditions when subjects can successfully subclassify test stimuli (Panel A) than when they can not (Panel B).

afford subclassification. Word frequency is the only member of this category from that list.

The other two variables, repetition and study time, are the paradigmatic examples of manipulations that do not routinely afford such subclassification. An encounter with a single test word reveals nothing about whether it was probably repeated or if it was probably studied for a long duration—other than through the evidence it yields for having been studied at all. Consistent with the explanation laid out here, these are the very variables for which the effects of discriminability on slope are less consistently observed.

To summarize, manipulations that encourage easy allotment of test items into discriminability classes are likely to promote lesser criterion variability and are thus less likely to mask the underlying decrease in the slope of the isosensitivity function generated by increasing stimulus variability.

Slope invariance with manipulations of bias. Another recent important result that is somewhat vexing from the standpoint of TSD is the lack of invariance in the shape of the isosensitivity function when estimated under different biasing or payoff conditions (Balakrishnan, 1998a; Van Zandt, 2000). This failure has led theorists to question some of the basic tenets of TSD, such as the assumption that confidence ratings are scaled from the evidence axis (Van Zandt, 2000) or, even more drastically, that stimulus distributions are not invariant with manipulations of bias (Balakrishnan, 1998b, 1999). Both suggestions do serious violence to the application of TSD to psychological tasks and to rating tasks in particular, but several theorists have defended the honor of the

venerable theory (Rotello & Macmillan, 2008; Treisman, 2002). Of particular interest here, a recent report by Mueller and Weidemann (2008) postulated criterial noise as a source of the failed invariance. Mueller and Weidemann demonstrated that criterial noise can account for the lack of invariance under a bias manipulation using their decision noise model, which is similar in spirit to (but quite different in application from) ND-TSD.

ND-TSD can also explain such effects quite simply. Figure 6 shows the joint effects of stimulus and criterial variability on the slope of the isosensitivity function. A manipulation of bias is presumed not to affect either the location or the shape of the evidence distributions (cf. Balakrishnan, 1999) and should consequently have no effect on slope. The predictions of TSD are indicated by the darkest (bottom) line; any point on that line is a potential slope value, and it should not change with a manipulation of bias. However, if criterion variability changes with bias, then the slope of the function can vary along a contour of constant stimulus variance, such as shown by Box A. Such an interpretation presumes that criterial variance itself varies with the bias manipulation; why might this be?

First, criterion variance might increase with increasing distance from an unbiased criterion. This could be true because placing criteria in such locations is uncommon or unfamiliar or simply because the location value is represented as a distance from the intersection of the distributions. A magnitude representation of distance would exhibit scalar variability and thus imply greater criterion variance with more biased criterion locations. The model of Mueller and Weidemann (2008) achieves this effect by imposing greater variability on peripheral confidence criteria than on the central yes–no criterion; such a mechanism is neither included in nor precluded by ND-TSD. Similarly, criteria may exhibit scalar variability with increasing distance from the mean of the noise distribution. This assumption is supported somewhat by results that indicate that criterion noise increases with stimulus range in absolute identification tasks (Nosofsky, 1983).

In sum, if the variance of criteria scales with the magnitude of those criteria, manipulations of bias may be incorrectly interpreted as reflecting changes in the stimulus distributions. This does not reflect a fundamental failing of TSD but rather reveals conditions in which ND-TSD is necessary to explain the effects of decision noise on estimated isosensitivity functions.

Deterministic Versus Probabilistic Response Criteria

In earlier flashpoints over decision rules in choice tasks, some theorists suggested that the rule may be probabilistic rather than deterministic in form (e.g., Luce, 1959; Nachmias & Kocher, 1970; Parks, 1966; Thomas & Legge, 1970). From the perspective of TSD, the evidence value is compared with a criterion value, and a decision is made based on their ordering. This strategy leads to optimal performance, in terms of either payoff maximization or maximal number of correct responses, when that criterion is based on the likelihood ratio (Green & Swets, 1966). Regardless of how the criterion is placed and whether it is optimal or not, this is a deterministic response rule and differs from a probabilistic response rule, by which the value of the likelihood ratio or a transformation thereof is continuously related to the probability of a particular response.

There was a tremendous amount of research devoted to the resolution of this question in the 1960s and 1970s, in part because TSD made such a forceful claim that the rule is deterministic. A convincing answer was not apparent, however: The strong implications of static criteria were rejected by the data reviewed above, including sequential dependencies and changes in the slope of the isosensitivity function. Improvements in sensitivity over the course of individual tasks (e.g., Gundy, 1961; Zwislocki, Marie, Feldman, & Rubin, 1958) also suggested the possibility of increasingly optimal or perhaps decreasingly variable criteria. In some tasks, the prediction of a binary cutoff in response probability that followed from deterministic theories was confirmed (Kubovy, Rapaport, & Tversky, 1971), and in other tasks, that prediction was disconfirmed (Lee & Janke, 1965; Lee & Zentall, 1966). In still others, data fell in a range that was not naturally predicted by either a binary cutoff or one of the probabilistic viewpoints reviewed below (Lee & Janke, 1964). Cutoffs appeared to be steeper when discriminability was greater (Lee & Zentall, 1966), suggesting that subjects may employ cutoffs within a range of evidence and use alternate strategies when the evidence less clearly favors one choice or the other (Parducci & Sandusky, 1965; Sandusky, 1971; Ward, 1973; Ward & Lockhead, 1971).

The evidence in favor of probabilistic models is mixed as well. The most general prediction of probabilistic models of decision making is that the probability of an endorsement varies with the evidence in favor of the presence of the to-be-endorsed stimulus. Whereas it is optimal to respond "old" to a recognition test stimulus when the evidence in favor of that stimulus actually having been studied outweighs the evidence that it was not (assuming equal priors and payoffs), probabilistic models suggest that the weight of that evidence determines the probability of an "old" response. Evidence from individual response functions in tasks that minimized sources of variability revealed sharp cutoffs (Kubovy et al., 1971). Simple probabilistic models failed to account for that result but accounted well for performance in other tasks (Schoeffler, 1965), including recognition memory (Parks, 1966) and a wide variety of higher level categorization tasks (Erev, 1998).

A partial reconciliation of these views came in the form of deterministic dynamic-criterion models (Biderman, Dorfman, & Simpson, 1975; Dorfman, 1973; Dorfman & Biderman, 1971; Kac, 1962, 1969), in which the criterion varied systematically from trial to trial on the basis of the stimulus, response, and outcome. These models outperformed models with static criteria (Dorfman & Biderman, 1971; Larkin, 1971) but did not account for a relatively large amount of apparently nonsystematic variability (Dorfman, Saslow, & Simpson, 1975). Similar models were proposed with probabilistic responding (Larkin, 1971; Thomas, 1973) but were never tested against dynamic-criterion models with deterministic responding.

Probability matching and base-rate manipulations. Many of the dynamic-criterion models made the prediction that responding would exhibit *probability matching* (or micromatching; Lee, 1963); that is, the probability of a positive response would asymptotically equal the a priori probability of a to-be-endorsed stimulus being presented (Creelman & Donaldson, 1968; Parks, 1966; Thomas & Legge, 1970). Such theories also met with mixed results: Although there were situations in which probability matching appeared to hold (e.g., Lee, 1971; Parks, 1966), time-series

analysis revealed overly conservative response frequencies (to be reviewed in greater detail below) and poor fits to individual subjects (Dusoir, 1974; M. F. Norman, 1971). Kubovy and Healy (1977) even concluded that dynamic-criterion models that employed error correction were mostly doomed to fail because, empirically, subjects appeared to shift criteria after both correct and incorrect responses, an effect that was inconsistent with the majority of models. They also claimed that models of the additive-operator type—in which the direction of criterion change following a correct response combination is predicted to be constant—were wrong because subjects appeared to be willing to shift their criterion in either direction, depending on the exact circumstances. Here, we have explicitly avoided theorizing about the nature of systematic changes in criterion so as to be able to more fully examine the role of nonsystematic noise on the response function and thus on recognition performance. Yet it can be shown that criterial noise naturally and simply leads to conservative shifts of criteria in response to manipulations of base rates of signal and noise events.

Conservatism. An important result in tasks in which base rates are manipulated is the excessively conservative response of criteria to manipulations of the base rates of events.[6] Overall, experiments have revealed mixed effects of base-rate manipulations: Although, in some tasks, subjects appear acutely sensitive to prior probabilities (Kubovy & Healy, 1977; Swets, Tanner, & Birdsall, 1961), even in recognition memory (A. F. Healy & Kubovy, 1978), those shifts typically are lesser in magnitude than predicted either under an optimal deterministic decision rule (Green & Swets, 1966) or under the more conservative prediction of probability matching (Thomas, 1975). Other data suggested that subjects in recognition memory experiments did not modulate their criteria at all when the base rates were shifted across blocks (A. F. Healy & Jones, 1975; A. F. Healy & Kubovy, 1977).

The general conservatism of criterion placement has been attributed to, variously, unwillingness to abandon sensory (or mnemonic) evidence in favor of base rates (Green & Swets, 1966), failure to appreciate the proper form of the evidence distributions (Kubovy, 1977), inaccurate estimation of prior probabilities (Galanter, 1974; Ulehla, 1966), or probability matching (although this latter view was eventually rejected by the data discussed in the previous section). In this next subsection, we show that either probability matching or optimality perspectives can predict conservatism when criterial variability is explicitly accounted for. Likewise, we show that criterial variability can mimic probabilistic response selection.

Response functions, conservatism, and manipulations of base rates. The most fundamental effect of the addition of criterial noise is to change the shape of the response function—that is, the function relating evidence to response. Here, we consider the form of response functions in the presence of criterial variability and evaluate the exact effect of that variability on the specific predictions of optimality views (Green & Swets, 1966) and probability matching (Parks, 1966). We show that (a) probabilistic response functions are not to be distinguished from deterministic functions with criterial variability and that (b) conservatism in criterion shifts in response to manipulations of base rates is a natural consequence of criterial variability (for more general arguments about mimicry between deterministic and probabilistic response functions, see Marley, 1992; Townsend & Landon, 1982). The goal of these claims is to show how criterial variability can increase the range of results that fall within the explanatory purview of TSD and to demonstrate why previously evaluated benchmarks for the rejection of deterministic models may be inappropriate. Specifically, ND-TSD naturally accounts for (apparently) suboptimal response probabilities in response to manipulations of base rate. It does so successfully because, as shown below, a deterministic response rule in the presence of criterial noise can perfectly mimic a probabilistic rule (for similar demonstrations, see Ashby & Maddox, 1993; Marley, 1992; Townsend & Landon, 1982).

The deterministic response rule is to endorse a stimulus as "old" if the subjective evidence value (E) surpasses a criterion value c:

$$p(\text{yes}) = p(E \geq c).$$

Treating c as an instance of the previously defined random variable for criterion, the response function conditional upon E is

$$p(\text{yes}|E) = \Phi\left(\frac{E - \mu_c}{\sigma_C}\right). \tag{10}$$

Example response functions are shown in Figure 8A, in which increasingly bright lines indicate increasingly variable criteria. The function is, of course, simply the cumulative normal distribution of which the step function that is the traditional implication of TSD (shown in black) is the asymptotic form as $\sigma_c^2 \rightarrow 0$. This result is not surprising, but it is revealing, especially in comparison with Figure 8B, which depicts response functions for two purely probabilistic response rules. The first (darker) depicts Schoeffler's (1965) response rule, which is

$$p(\text{yes}|E) = \frac{\Phi\left(\frac{E - \mu_1}{\sigma_1^2}\right)}{1 - \Phi(E) + \Phi\left(\frac{E - \mu_1}{\sigma_1^2}\right)},$$

and the second (lighter) line depicts an even simpler rule relating the height of the signal distribution at E to the sum of the heights of the two distributions at E, or

$$p(\text{yes}|E) = \frac{\varphi\left(\frac{E - \mu_1}{\sigma_1^2}\right)}{\varphi(E) + \varphi\left(\frac{E - \mu_1}{\sigma_1^2}\right)},$$

in which φ indicates the normal probability density function. In each of these cases, the resultant response function is also a cumulative normal distribution, thus showing that criterial variability can make a deterministic response rule perfectly mimic a probabilistic one. To be fair, the rules chosen here are simple ones, and simplifying assumptions have been made with respect to the evidence distributions (with the latter rule, the evidence distributions have been set to be of equal variance). It is not our claim that there are not probabilistic rules that may be differentiated from deterministic rules with criterial noise, nor that there are no cir-

[6] Note that *conservative* in this context refers to a suboptimal magnitude of criterion shift with respect to changing base rates, not to a conservative (as opposed to liberal) criterion placement.

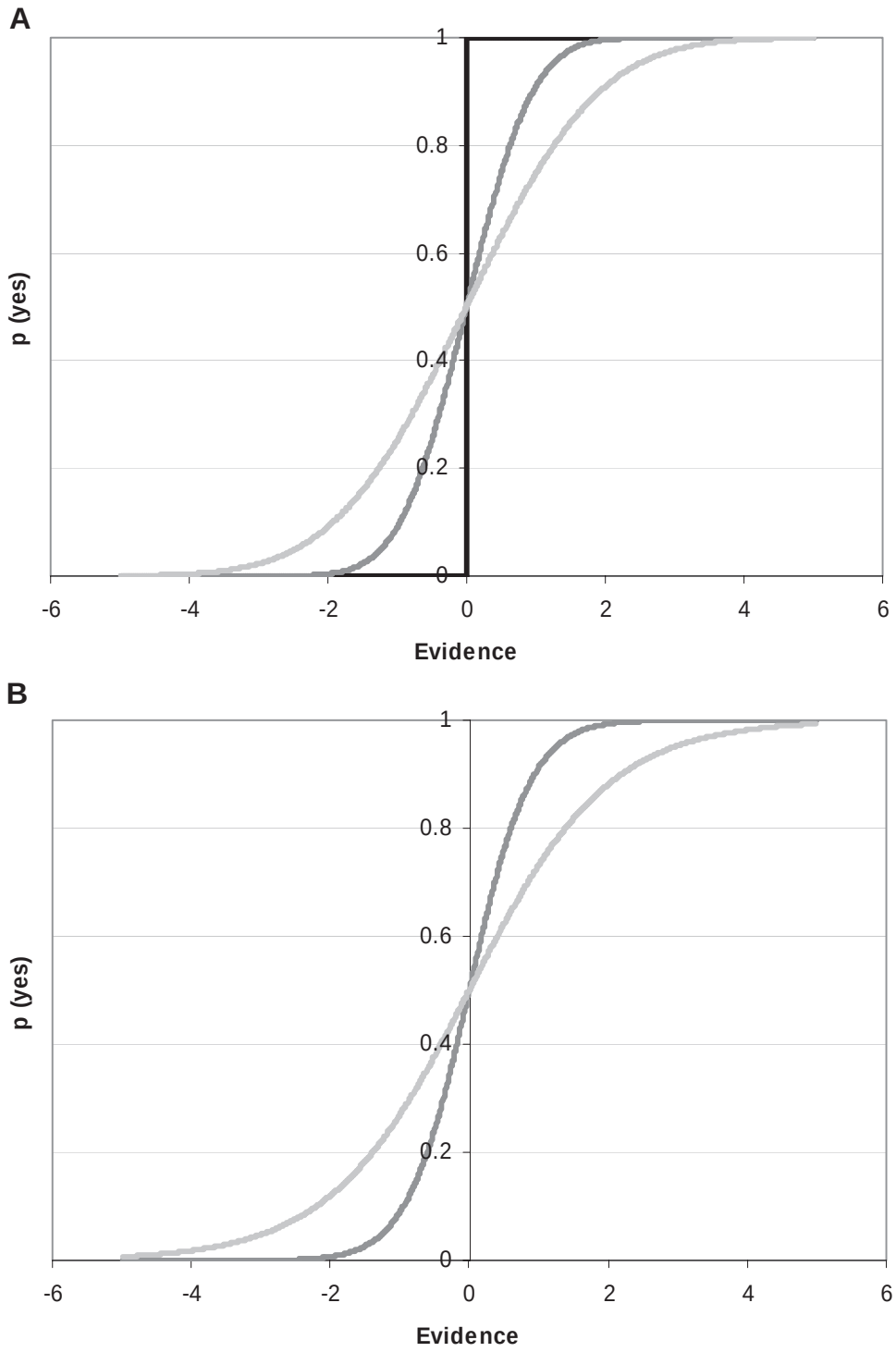

Figure 8. Response functions for deterministic response rules (Panel A) and probabilistic response rules (Panel B). In Panel A, increasingly light lines indicate increasing criterial noise. Note that it is a step function when the criterion is nonvariable. In Panel B, the two functions represent two different response rules (see text for details). In all cases, the criterion is set at 0.

cumstances under which even these rules can be differentiated from one another. Rather, it is to demonstrate that a parameter governing criterial variability can produce a range of response functions, including ones that perfectly replicate the predictions of probabilistic rules. This result provides a new perspective on the phenomenon of conservatism seen in criterion setting, as we review below.

Conservatism in response to base-rate manipulations. The conservatism seen in responses to manipulations of base rate has been hypothesized to reflect either suboptimal criterion placement or a failure to accurately estimate the parameters of the decision regime, including the probability distributions or the actual base rates themselves. Here, we show that conservatism is a natural consequence of criterial variability and arises with both optimal criterion placement and probability matching strategies.

Optimal criteria for base-rate manipulations. Green and Swets (1966) showed that the optimal bias can be defined purely in terms of the stimulus base rates:

$$\beta_{\text{optimal}} = \frac{p\,(\text{noise})}{p\,(\text{signal})}.$$

When the evidence distributions are not of equal variance, an optimal bias leads to two criteria. This fact is reflected in the nonmonotonicity at the margins of the isosensitivity function or, equivalently, by the nonmonotonic relationship between evidence and the likelihood ratio throughout the scale. In any case, this issue falls outside the purview of our current discussion and need not concern us here. The effect of criterial variability can be amply demonstrated under the equal-variance assumption.

In the equal-variance case, the optimal criterion placement is a function of the optimal bias and the distance between the distributions:

$$C_{\text{optimal}} = \frac{\log \beta_{\text{optimal}} + \frac{1}{2} d'^2}{d'},$$

in which d' represents the distance between the evidence distributions scaled by their common standard deviation.

Imagine that subjects place their criterion optimally according to this analysis but fail to account for the presence and consequence of criterial noise. To evaluate that effect, we must consider first how criterial variability affects d'. To do so, remember that $d' = z\text{HR} - z\text{FAR}$. Substituting terms from Equation 1b and setting $\sigma_1^2 = \sigma_0^2 = 1$,

$$d'_{\text{noisy}} = \frac{\mu_1 - \mu_C}{\sqrt{1 + \sigma_C^2}} - \frac{-\mu_C}{\sqrt{1 + \sigma_C^2}} = \frac{\mu_1}{\sqrt{1 + \sigma_C^2}}, \qquad (11)$$

where d'_{noisy} indicates d' under conditions of criterial variability.

This relationship indicates that d' will be overestimated in computing optimal criterion placement and that this overestimation will worsen with increasing criterial noise. What effect does this have on the overall rate of positive responding? That relationship is shown in Figure 9A, which plots the deviation of overall "yes" rate from the predicted rate of a semi-ideal decision maker—that is, one that is ideal except insofar as it fails to appreciate its own criterial noise. These values were computed by assessing the rate of positive responding (for to-be-endorsed and to-be-rejected stimuli) at the semi-ideal criterion for varying base rates (for $d' = 1$) and then comparing that value with the rate of responding with added criterial noise. As noted by Thomas and Legge (1970), the effect of criterial variability is to lead to the appearance of nonoptimal criterion placement. The employed criterion is optimal from the perspective of the information available in the task but nonoptimal in that it fails to account for its own variability. The net effect is that low signal probabilities lead to a nonoptimally high rate of responding and high signal probabilities lead to a nonoptimally low rate of responding. This result is the hallmark of conservatism.

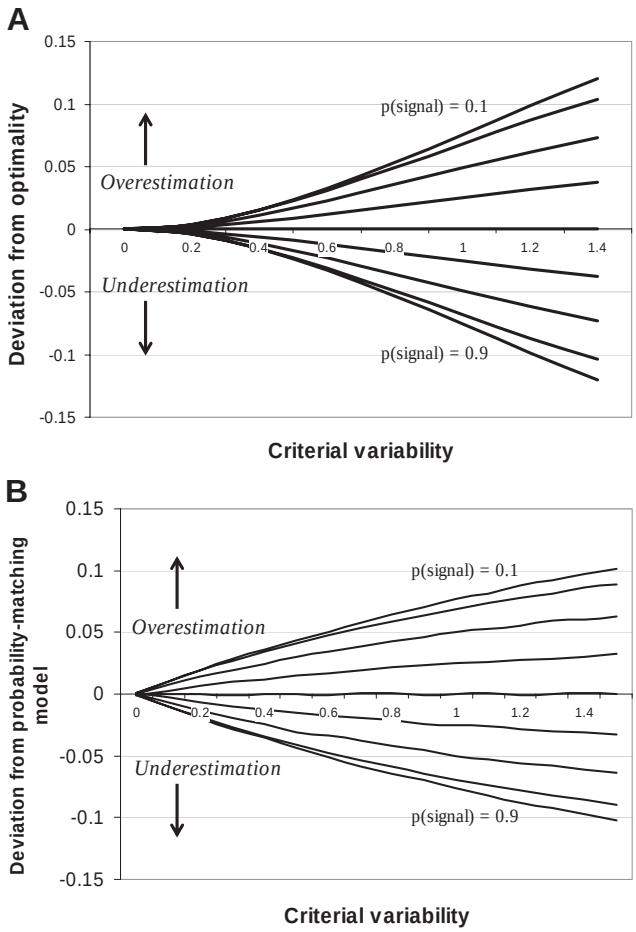

Figure 9. A demonstration of conservatism as a function of base-rate manipulations. Panel A shows deviation from optimal responding; Panel B shows deviation from probability matching.

Probability matching and criterial noise. According to the probability matching view, subjects aim to respond positively at the same rate as the positive signal is presented. Probability matching often predicts more conservative response behavior than does the optimality view presented above (Thomas & Legge, 1970). Let P_1 be the proportion of signal trials and thus also the desired rate of positive responding (R_1). Then,

$$R_1 = P_1(\text{HR}) + (1 - P_1)(\text{FAR})$$
$$= P_1 * \Phi\left(\frac{\mu_1 - \mu_C}{\sigma_1^2 + \sigma_C^2}\right) + (1 - P_1) * \Phi\left(\frac{-\mu_C}{1 + \sigma_C^2}\right). \qquad (12)$$

Thomas and Legge (1970) pointed out that this function is not an isosensitivity function but rather an isocriterion function: It describes the relationship between HR and FAR and that relationship's invariance with R_I as sensitivity varies. Thus, like the case above, we must assume a particular level of sensitivity to derive values for the HR and FAR. In addition, the relationship between R_I and μ_C is complex because of the integral over the normal distribution. In the simulation that follows, we selected the value for μ_C that minimized the deviation of the right-hand portion of Equation 12 from P_1, assuming values of 1 for μ_1 and σ_1. Deviation from this model was estimated by simulating 1,000,000 trials for each signal base rate from 0.1 to 0.9 (by steps of 0.1) and adding a variable amount of criterial noise on each trial. The results, shown in Figure 9B, indicate an effect similar to what was seen in the previous case: Increasing criterial noise leads to increased conservatism.

Criterion noise within a deterministic decision framework (TSD) can thus account for results that have been proposed to reveal probabilistic responding. These demonstrations in and of themselves do not reveal the superiority of deterministic theories, but they do suggest that such data are not decisive for either viewpoint and, in doing so, call attention to the large additional body of evidence in support of deterministic decision theories.

The Effects of Aging on Recognition

Many of the current battles over the nature of the information that subserves recognition decisions are waged using data that compare age groups. For example, it has been proposed that older subjects specifically lack recollective ability (Jacoby, 1991; Mandler, 1980) but enjoy normal levels of familiarity. Evidence for this two-component theory of recognition comes from age-related dissociations in performance as well as differences between younger and older subjects in the shape of the isosensitivity function (Howard et al., 2006; Yonelinas, 2002). However, the role of criterial variability has never been considered.

Two general sources of differences between age groups in criterial maintenance are possible. First, those mechanisms and strategies that govern the evolution of criterion placement over time may differ between young and older subjects, perhaps leading to differences in variability of that placement over the course of the experiment. Such a finding would be fascinating in that it would provide an example of how higher level cognitive strategic differences play out in terms of performance on very basic tests of memory (cf. Benjamin & Ross, 2008). Alternatively, it might be the case that maintenance of criterion is simply a noisier process in older people—perhaps attributable to one of very problems in older people that it can be confused with, namely, memory (Kester, Benjamin, Castel, & Craik, 2002)—and that recognition suffers as a result.

Empirically, the results are as one might expect if older adults exhibit greater criterion variability. The slope of the isosensitivity function is greater for older than younger subjects on tasks of word recognition (Kapucu, Rotello, Ready, & Seidl, 2008), picture recognition (Howard et al., 2006), and associative recognition (M. R. Healy et al., 2005). The wide variety of materials across which this age-related effect obtains is suggestive of a quite general effect of aging on criterion maintenance, rather than a strategic difference between the age groups. These studies have not attempted to separate the effects of criterion and stimulus variability, and these results are consistent with, but not uniquely supportive of, greater decision noise in older adults. Future work is necessary to isolate these effects within older subjects.

Changes in Sensitivity With Time

TSD is often used to evaluate whether fatigue affects performance on a detection task over time (Galinsky, Rosa, Warm, & Dember, 1993; cf. Dobbins, Tiedmann, & Skordahl, 1961) or, conversely, whether improvements in sensitivity are evident with increasing practice (Gundy, 1961; Trehub, Schneider, Thorpe, & Judge, 1991; Zwislocki et al., 1958). Traditional interpretations of such effects attribute fatigue-related decrements to increasing stimulus noise and practice-related improvement to increasing criterion optimization, but such dramatically differing interpretations of these related effects are not compelled by the data. They reflect a tacit but intuitive belief that maintenance of criteria is not demanding and thus not subject to fatigue. The purely perceptual part of detection tasks is assumed to be similarly undemanding and thus not likely to show much improvement with practice.

Consideration of criterial variability provides an alternate theoretical rationale that can unite these findings: Decrements arise with time when fatigue increases the difficulty of criterial maintenance, and improvements arise when practice decreases the effects of noise on criterion localization. Such a statement should not be confused with an articulated psychological theory of such effects, but it is an alternative theoretical mechanism that such theories might profitably take advantage of when substantively addressing these and related results.

The Slope of the Isosensitivity Function for Remember–Know Tasks

There is currently a vigorous debate over judgments that subjects provide about the phenomenological nature of their recognition judgments and whether those judgments validly represent different sources of evidence (Gardiner & Gregg, 1997; Gardiner, Richardson-Klavehn, & Ramponi, 1998) or two criteria applied to a single continuous evidence dimension (Benjamin, 2005; Donaldson, 1996; Dunn, 2004; Hirshman & Master, 1997; Wixted & Stretch, 2004). The latter view is consistent with the received version of unidimensional TSD with multiple criteria, as in the ratings task discussed previously at length, whereas the former view specifies additional sources of evidence beyond those captured in a single evidence dimension. Which view is correct is a major theoretical debate for theorists of recognition memory, and whether these phenomenological judgments of "remember" and "know" status indicate multiple states or multiple criteria has become a major front in that battle. That debate is peripheral to the present work and is not reviewed here. We do consider how criterial variability might influence interpretation of data relevant to that debate, however.

Some authors have cited differences in the slope of the isosensitivity function estimated from remember–know judgments from the slope estimated from confidence ratings as evidence against the unidimensional view of remember–know judgments (Rotello,

Macmillan, & Reeder, 2004), whereas others have disputed this claim (Wixted & Stretch, 2004). In a large meta-analysis, Rotello et al. (2004) examined slopes for isosensitivity functions relating remember responses to overall rates of positive responses and found a greater slope for such isosensitivity functions. Wixted and Stretch (2004) explained this result thusly: "The evidence suggests that the location of the remember criterion exhibits item-to-item variability with respect to the confidence criteria . . . if the remember criterion varies from item to item, the slope of the [isosensitivity function] would increase accordingly" (p. 627).

Although not described in the same framework that we provide here, the astute reader will recognize a claim of criterial variability analogous to our earlier discussion. If the judgments in the remember–know paradigm are subject to greater variability than the judgments in a confidence-rating procedure, then the slope of the isosensitivity function will be closer to 1 for the remember–know function than for the confidence function. If the slope of the confidence function is less than 1, as it typically is, then the additional criterial variability associated with remember–know judgments will increase the slope of the function. This is exactly the result reported by Rotello et al. (2004).

This interpretation is further borne out by recent studies that empirically assessed the variability in the location of the remember–know criterion. Recent studies that compared models of remember–know judgments with and without an allowance for criterion variability for the criterion lying between "know" and "remember" judgments revealed superior performance by the models with nonzero criterion variability (Dougal & Rotello, 2007; Kapucu et al., 2008). The heady controversy underlying the use of the remember–know procedure may thus reflect the consequences of unconsidered noise in the decision process.

Detection Theory and Criterial Noise

This article has questioned a very basic assumption of the TSD—that the criterion value is a stable, stationary value. We have forwarded theoretical arguments based on the psychological burden of maintaining criteria and reviewed empirical evidence that suggests the presence of criterial noise, including comparisons of different procedures for estimating isosensitivity functions and systematic effects of experimental manipulations on criteria, and a long-standing debate over whether the response rule is probabilistic or deterministic. Criterial noise makes these two candidates indistinguishable and naturally accounts for the conservatism in response shifts that is ubiquitous in experiments that manipulate signal base rates. In addition, we have argued that the isosensitivity function can be used to test theories of recognition only if criterial variability is presumed to be negligible.

In the second half of the article, we have used the task of ensemble recognition to tease apart the effects of criterial and stimulus noise and shown that criterial noise can be quite substantial. Given the empirical variability in estimates of slopes of isosensitivity functions across conditions (Swets, 1986a) and the lack of a strong theory that naturally accounts for such inconsistency, it may be useful to consider criterion noise as a meaningful contributor to the shape of the isosensitivity function and to detection, discrimination, and recognition more generally.

We have considered at some length the psychological implications of this claim. The effects of learning on detection, discrimination, and recognition tasks have always been interpreted in terms of shifting evidence distributions. Primarily, distributions are thought to overlap less under conditions of superior memory, but hypotheses regarding the relationship of their shape to performance have also recently been discussed (DeCarlo, 2002; Hilford et al., 2002). The specifics of that shape have even been used to test the assumptions of competing models of the nature of recognition judgments (Heathcote, 2003; Yonelinas, 1999). Here, we have argued that learning may also influence the variability of criteria and that superior performance may in part reflect greater criterion stability. This explanation does not deemphasize the role of encoding and retention of stimuli as a basis for recognition performance, but it allows for task-relevant expertise over the course of the test to play an additional role.

Finally, we have reviewed a set of problems that the postulate of criterion noise might help provide new solutions for, including the inconsistency of manipulations of learning on the slope of the isosensitivity function, discrepancies between procedures used to estimate such functions, the effects of prior odds on shifts in response policy, the nature of remember–know judgments in recognition, the effects of fatigue on judgment tasks of vigilance, and the effects of aging on recognition. This is a small subset of areas in which decision noise is relevant, but it illustrates the dilemma: Accurate separation of the mnemonic aspects of recognition from the decision components of recognition relies on valid assumptions about the reliability, as well as the general nature, of the decision process. We have provided evidence that variability in this process is important, is apparent, and undermines attempts to use TSD as a general means of evaluating models of recognition. ND-TSD reconciles the powerful theoretical machinery of TSD with realistic assumptions about the fallibility of the decision process.

References

Aggleton, J. P., & Shaw, C. (1996). Amnesia and recognition memory: A re-analysis of psychometric data. *Neuropsychologia, 34,* 51–62.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second International Symposium on Information Theory* (pp. 267–281). Budapest, Hungary: Akademiai Kiado.

Arndt, J., & Reder, L. M. (2002). Word frequency and receiver operating characteristic curves in recognition memory: Evidence for a dual-process interpretation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 830–842.

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

Atkinson, R. C., Carterette, E. C., & Kinchla, R. A. (1964). The effect of information feedback upon psychophysical judgments. *Psychonomic Science, 1,* 83–84.

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. (1998a). Measures and interpretations of vigilance performance: Evidence against the detection criterion. *Human Factors, 40,* 601–623.

Balakrishnan, J. D. (1998b). Some more sensitive measures of sensitivity and response bias. *Psychological Methods, 3,* 68–90.

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.

Balota, D. A., Cortese, M. J., Hutchison, K. A., Neely, J. H., Nelson, D., Simpson, G. B., &, & Treiman, R. (2002). *The English Lexicon Project: A Web-based repository of descriptive and behavioral measures for 40,481 English words and nonwords.* Retrieved from the Washington University Web-site: http://elexicon.wustl.edu/

Banks, W. P. (1970). Signal detection theory and human memory. *Psychological Bulletin, 74,* 81–99.

Banks, W. P. (2000). Recognition and source memory as multivariate decision processes. *Psychological Science, 11,* 267–273.

Bayley, P. J., Wixted, J. T., Hopkins, R. O., & Squire, L. R. (2008). Yes/no recognition, forced-choice recognition, and the human hippocampus. *Journal of Cognitive Neuroscience, 20,* 505–512.

Benjamin, A. S. (2001). On the dual effects of repetition on false recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 941–947.

Benjamin, A. S. (2003). Predicting and postdicting the effects of word frequency on memory. *Memory & Cognition, 31,* 297–305.

Benjamin, A. S. (2005). Recognition memory and introspective remember/know judgments: Evidence for the influence of distractor plausibility on "remembering" and a caution about purportedly nonparametric measures. *Memory & Cognition, 33,* 261–269.

Benjamin, A. S., & Bawa, S. (2004). Distractor plausibility and criterion placement in recognition. *Journal of Memory and Language, 51,* 159–172.

Benjamin, A. S., Bjork, R. A., & Hirshman, E. (1998). Predicting the future and reconstructing the past: A Bayesian characterization of the utility of subjective fluency. *Acta Psychologica, 98,* 267–290.

Benjamin, A. S., Lee, J., & Diaz, M. (2008). *Criterion noise in recognition memory judgments.* Manuscript in preparation.

Benjamin, A. S., & Ross, B. H. (Eds.). (2008). *The psychology of learning and motivation: Vol. 48. Skill and strategy in memory use.* San Diego, CA: Academic Press.

Benjamin, A. S., Wee, S., & Roberts, B. W. (2008). [The relationship between narcissism, overclaiming of knowledge, and false-alarm rates in recognition]. Unpublished raw data.

Biderman, M., Dorfman, D. D., & Simpson, J. C. (1975). A learning model for signal detection theory-temporal invariance of learning parameters. *Bulletin of the Psychonomic Society, 6,* 329–330.

Birdsall, T. G. (1966). The theory of signal detectability: ROC curves and their character. *Dissertation Abstracts International, 28,* 1B.

Blackwell, H. (1953). *Psychophysical thresholds: Experimental studies of methods of measurement* (University of Michigan Engineering Research Institute Bulletin No. 36). Ann Arbor: University of Michigan, Engineering Research Institute.

Bonnel, A.-M., & Miller, J. (1994). Attentional effects on concurrent psychophysical discriminations: Investigations of a sample-size model. *Perception & Psychophysics, 55,* 162–179.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10,* 433–436.

Brown, J., Lewis, V. J., & Monk, A. F. (1977). Memorability, word frequency and negative recognition. *Quarterly Journal of Experimental Psychology, 29,* 461–473.

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

Brown, S., Steyvers, M., & Hemmer, P. (2007). Modeling experimentally induced strategy shifts. *Psychological Science, 18,* 40–45.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research, 33,* 261–304.

Creelman, C. D., & Donaldson, W. (1968). ROC curves for discrimination of linear extent. *Journal of Experimental Psychology, 77,* 514–516.

Criss, A. H. (2006). The consequences of differentiation in episodic memory: Similarity and the strength based mirror effect. *Journal of Memory and Language, 55,* 461–478.

DeCarlo, L. T. (2002). Signal detection theory with finite mixture distributions: Theoretical developments with applications to recognition memory. *Psychological Review, 109,* 710–721.

Deffenbacher, K. A., Leu, J. R., & Brown, E. L. (1981). Memory for faces: Testing method, encoding strategy, and confidence. *American Journal of Psychology, 94,* 13–26.

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review, 108,* 452–478.

Dobbins, D. A., Tiedmann, J. G., & Skordahl, D. M. (1961). *Field study of vigilance under highway driving conditions* (Technical Research Note No. 118). Arlington, VA: U.S. Army Personnel Research Office, Office of the Chief of Research and Development.

Donaldson, W. (1996). The role of decision processes in remembering and knowing. *Memory & Cognition, 24,* 523–533.

Dorfman, D. D. (1973). The likelihood function of additive learning models: Sufficient conditions for strict log-concavity and uniqueness of maximum. *Journal of Mathematical Psychology, 10,* 73–85.

Dorfman, D. D., & Biderman, M. (1971). A learning model for a continuum of sensory states. *Journal of Mathematical Psychology, 8,* 264–284.

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.

Dougal, S., & Rotello, C. M. (2007). "Remembering" emotional words is based on response bias, not recollection. *Psychonomic Bulletin & Review, 14,* 423–429.

Dunn, J. C. (2004). Remember–know: A matter of confidence. *Psychological Review, 111,* 524–542.

Durlach, N. I., & Braida, L. D. (1969). Intensity perception: I. Preliminary theory of intensity resolution. *Journal of the Acoustical Society of America, 46,* 372–383.

Dusoir, A. E. (1974). Thomas and Legge's matching hypothesis for detection and recognition tasks: Two tests. *Perception & Psychophysics, 16,* 466–470.

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

Egan, J. P., Greenberg, G. Z., & Schulman, A. I. (1961). Operating characteristics, signal detectability, and the method of free response. *Journal of the Acoustical Society of America, 33,* 993–1007.

Egan, J. P., Schulman, A. I., & Greenberg, G. Z. (1959). Operating characteristics determined by binary decisions and by ratings. *Journal of the Acoustical Society of America, 31,* 768–773.

Eich, J. M. (1982). A composite holographic associative recall model. *Psychological Review, 89,* 627–661.

Elam, L. E. (1991). *Variance of memory-strength distributions and the list-length effect.* Unpublished bachelor's thesis, University of Oklahoma, Norman.

Emmerich, D. S. (1968). Receiver-operating characteristics determined under several interaural conditions of listening. *Journal of the Acoustical Society of America, 43,* 298–307.

Erev, I. (1998). Signal detection by human observers: A cutoff reinforcement learning model of categorization decisions under uncertainty. *Psychological Review, 105,* 280–298.

Estes, W. K., & Maddox, W. T. (2005). Risks of drawing inferences about cognitive processes from model fits to individual versus average performance. *Psychonomic Bulletin & Review, 12,* 403–408.

Fechner, G. T. (1860). *Elemente der Psychophysik* [Elements of psychophysics]. Leipzig, Germany: Breitkopf und Härtel.

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

Freed, D. M., Corkin, S., & Cohen, N. J. (1987). Forgetting in H. M.: A second look. *Neuropsychologia, 25,* 461–471.

Friedman, M. P., Carterette, E. C., Nakatani, L., & Ahumada, A. (1968). Comparisons of some learning models for response bias in signal detection. *Perception & Psychophysics, 3,* 5–11.

Galanter, E. (1974). Psychological decision mechanisms and perception. In E. C. Carterette & M. P. Friedman (Eds.), *Handbook of perception: Vol. 2. Psychophysical judgment and measurement* (pp. 85–126). New York: Academic Press.

Galinsky, T. L., Rosa, R. R., Warm, J. S., & Dember, W. N. (1993). Psychophysical determinants of stress in sustained attention. *Human Factors, 35,* 603–614.

Gardiner, J. M., & Gregg, V. H. (1997). Recognition memory with little or no remembering: Implications for a detection model. *Psychonomic Bulletin & Review, 4,* 474–479.

Gardiner, J. M., Richardson-Klavehn, A., & Ramponi, C. (1998). Limitations of the signal-detection model of the remember–know paradigm: A reply to Hirshman. *Consciousness & Cognition, 7,* 285–288.

Gilden, D. L. (1997). Fluctuations in the time required for elementary decisions. *Psychological Science, 8,* 296–301.

Gilden, D. L. (2001). Cognitive emissions of 1/f noise. *Psychological Review, 108,* 33–56.

Gilden, D. L., & Wilson, S. G. (1995). On the nature of streaks in signal detection. *Cognitive Psychology, 28,* 17–64.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91,* 1–67.

Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory & Cognition, 13,* 8–20.

Glanzer, M., & Adams, J. K. (1990). The mirror effect in recognition memory: Data and theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 5–16.

Glanzer, M., Adams, J. K., Iverson, G. J., & Kim, K. (1993). The regularities of recognition memory. *Psychological Review, 100,* 546–567.

Glanzer, M., Kim, K., Hilford, A., & Adams, J. K. (1999). Slope of the receiver-operating characteristic in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 500–513.

Gravetter, F., & Lockhead, G. R. (1973). Criterial range as a frame of reference for stimulus judgment. *Psychological Review, 80,* 203–216.

Green, D. M. (1964). General prediction relating yes–no and forced-choice results [Abstract]. *Journal of the Acoustical Society of America, 36,* 1042.

Green, D. M., & Moses, F. L. (1966). On the equivalence of two recognition measures of short-term memory. *Psychological Bulletin, 66,* 228–234.

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

Greene, R. (2007). Foxes, hedgehogs, and mirror effects: The role of general principles in memory research. In J. S. Nairne (Ed.), *The foundations of remembering: Essays in honor of Henry L. Roediger III* (pp. 53–66). New York: Psychology Press.

Gronlund, S. D., & Elam, L. E. (1994). List-length effect: Recognition accuracy and variance of underlying distributions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1355–1369.

Gundy, R. F. (1961). Auditory detection of an unspecified signal. *Journal of the Acoustical Society of America, 33,* 1008–1012.

Healy, A. F., & Jones, C. (1975). Can subjects maintain a constant criterion in a memory task? *Memory & Cognition, 3,* 233–238.

Healy, A. F., & Kubovy, M. (1977). A comparison of recognition memory to numerical decision: How prior probabilities affect cutoff location. *Memory & Cognition, 5,* 3–9.

Healy, A. F., & Kubovy, M. (1978). The effects of payoffs and prior probabilities on indices of performance and cutoff location in recognition memory. *Memory & Cognition, 6,* 544–553.

Healy, M. R., Light, L. L., & Chung, C. (2005). Dual-process models of associative recognition in young and older adults: Evidence from receiver operating characteristics. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 768–788.

Heathcote, A. (2003). Item recognition memory and the receiver operating characteristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 1210–1230.

Higham, P. A., Perfect, T. J., & Bruno, D. (in press). Investigating strength and frequency effects in recognition memory using Type-2 Signal Detection Theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

Hilford, A., Glanzer, M., Kim, K., & DeCarlo, L. T. (2002). Regularities of source recognition: ROC analysis. *Journal of Experimental Psychology: General, 131,* 494–510.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93,* 411–428.

Hirshman, E. (1995). Decision processes in recognition memory: Criterion shifts and the list-strength paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 302–313.

Hirshman, E., & Hostetter, M. (2000). Using ROC curves to test models of recognition memory: The relationship between presentation duration and slope. *Memory & Cognition, 28,* 161–166.

Hirshman, E., & Master, S. (1997). Modeling the conscious correlates of recognition memory: Reflections on the remember–know paradigm. *Memory & Cognition, 25,* 345–351.

Hockley, W. E., & Niewiadomski, M. W. (2001). Interrupting recognition memory: Tests of a criterion-change account of the revelation effect. *Memory & Cognition, 29,* 1176–1184.

Howard, M. W., Bessette-Symons, B., Zhang, Y., & Hoyer, W. J. (2006). Aging selectively impairs recollection in recognition memory for pictures: Evidence from modeling and receiver operating characteristic curves. *Psychology and Aging, 21,* 96–106.

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

Humphreys, M., Pike, R., Bain, J. D., & Tehan, G. (1989). Global matching: A comparison of the SAM, Minerva II, Matrix, and TODAM models. *Journal of Mathematical Psychology, 33,* 36–67.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30,* 513–541.

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

Kac, M. (1969, November 7). Some mathematical models in science. *Science, 166,* 695–699.

Kapucu, A., Rotello, C. M., Ready, R. E., & Seidl, K. N. (2008). Response bias in "remembering" emotional stimuli: A new perspective on age differences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 703–711.

Kelley, R., & Wixted, J. T. (2001). On the nature of associative information in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 701–722.

Kester, J. D., Benjamin, A. S., Castel, A. D., & Craik, F. I. M. (2002). Memory in elderly people. In A. Baddeley, B. Wilson, & M. Kopelman (Eds.), *Handbook of memory disorders* (2nd ed., pp. 543–568). London: Wiley.

Khoe, W., Kroll, N. E. A., Yonelinas, A. P., Dobbins, I. G., & Knight, R. T. (2000). The contribution of recollection and familiarity to yes–no and forced-choice recognition tests in healthy subjects and amnesics. *Neuropsychologia, 38,* 1333–1341.

Kornbrot, D. E. (1980). Attention bands: Some implications for categorical judgment. *British Journal of Mathematical and Statistical Psychology, 33,* 1–16.

Krantz, D. H. (1969). Threshold theories of signal detection. *Psychological Review, 76,* 308–324.

Kroll, N. E. A., Yonelinas, A. P., Dobbins, I. G., & Frederick, C. M. (2002). Separating sensitivity from response bias: Implications of comparisons of yes–no and forced-choice tests for models and measures of recognition memory. *Journal of Experimental Psychology: General, 131,* 241–254.

Kubovy, M. (1977). A possible basis for conservatism in signal detection and probabilistic categorization tasks. *Perception & Psychophysics, 22,* 277–281.

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.

Kubovy, M., Rapoport, A., & Tversky, A. (1971). Deterministic vs. probabilistic strategies in detection. *Perception & Psychophysics, 9,* 427–429.

Lapsley Miller, J. A., Scurfield, B. K., Drga, V., Galvin, S. J., & Whitmore, J. (2002). Nonparametric relationships between single-interval and two-interval forced-choice tasks in the theory of signal detectability. *Journal of Mathematical Psychology, 46,* 383–417.

Larkin, W. (1971). Response mechanisms in detection experiments. *Journal of Experimental Psychology, 91,* 140–153.

Lee, W. (1963). Choosing among confusably distributed stimuli with specified likelihood ratios. *Perceptual and Motor Skills, 16,* 445–467.

Lee, W. (1971). Preference strength, expected value difference and expected regret ratio. *Psychological Bulletin, 75,* 186–191.

Lee, W., & Janke, M. (1964). Categorizing externally distributed stimulus samples for three continua. *Journal of Experimental Psychology, 68,* 376–382.

Lee, W., & Janke, M. (1965). Categorizing externally distributed stimulus samples for unequal molar probabilities. *Psychological Reports, 17,* 79–90.

Lee, W., & Zentall, T. R. (1966). Factorial effects in the categorization of externally distributed stimulus samples. *Perception & Psychophysics, 1,* 120–124.

Lockhart, R. S., & Murdock, B. B. (1970). Memory and the theory of signal detection. *Psychological Bulletin, 74,* 100–109.

Luce, R. D. (1959). On the possible psychophysical laws. *Psychological Review, 66,* 81–95.

Luce, R. D. (1963). A threshold theory for simple detection experiments. *Psychological Review, 70,* 61–79.

Luce, R. D., Nosofsky, R. M., Green, D. M., & Smith, A. F. (1982). The bow and sequential effects in absolute identification. *Perception & Psychophysics, 32,* 397–408.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ: Erlbaum.

Malmberg, K. J. (2002). On the form of ROCs constructed from confidence ratings. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 380–387.

Malmberg, K. J., & Xu, J. (2006). The influence of averaging and noisy decision strategies on the recognition memory ROC. *Psychonomic Bulletin & Review, 13,* 99–105.

Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review, 87,* 252–271.

Manns, J., Hopkins, R., Reed, J., Kitchener, E., & Squire, L. S. (2003). Recognition memory and the human hippocampus. *Neuron, 37,* 171–180.

Markowitz, J., & Swets, J. A. (1967). Factors affecting the slope of empirical ROC curves: Comparison of binary and rating responses. *Perception & Psychophysics, 2,* 91–97.

Marley, A. A. J. (1992). Developing and characterizing multidimensional Thurstone and Luce models for identification and preference. In G. F. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 299–333). Hillsdale, NJ: Erlbaum.

Matzen, L. E. & Benjamin, A. S. (in press). Remembering words not presented in sentences: How study context changes patterns of false memories. *Memory & Cognition.*

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review, 105,* 734–760.

McGill, W. J. (1957). Serial effects in auditory threshold judgments. *Journal of Experimental Psychology, 53,* 297–303.

McNicol, D. (1972). *A primer of signal detection theory.* Mahwah, NJ: Erlbaum.

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

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review, 89,* 609–626.

Myung, I. J., & Pitt, M. A. (2002). Mathematical modeling. In H. Pashler & J. Wixted (Eds.), *Stevens' handbook of experimental psychology: Vol. 4. Methodology in experimental psychology* (3rd ed., pp. 429–460). London: Wiley.

Nachmias, J., & Kocher, E. C. (1970). Visual detection and discrimination of luminance increments. *Journal of the Optical Society of America, 60,* 382–389.

Nachmias, J., & Steinman, R. M. (1963). Study of absolute visual detection by the rating-scale method. *Journal of the Optical Society of America, 53,* 1206–1213.

Norman, D. A., & Wickelgren, W. A. (1969). Strength theory of decision rules and latency in retrieval from short-term memory. *Journal of Mathematical Psychology, 6,* 192–208.

Norman, M. F. (1971). Statistical inference with dependent observations: Extensions of classical procedures. *Journal of Mathematical Psychology, 8,* 444–451.

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.

Ogilvie, J. C., & Creelman, C. D. (1968). Maximum-likelihood estimation of receiver operating characteristic curve parameters. *Journal of Mathematical Psychology, 5,* 377–391.

Parducci, A. (1984). Perceptual and judgmental relativity. In V. Sarris & A. Parducci (Eds.), *Perspectives in psychological experimentation* (pp. 135–149). Hillsdale, NJ: Erlbaum.

Parducci, A., & Sandusky, A. (1965). Distribution and sequence effects in judgment. *Journal of Experimental Psychology, 69,* 450–459.

Parks, T. E. (1966). Signal-detectability theory of recognition performance. *Psychological Review, 73,* 44–58.

Pelli, D. G. (1997). The Video Toolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10,* 437–442.

Peterson, W. W., Birdsall, T. G., & Fox, W. C. (1954). The theory of signal detectability. *Transactions of the IRE Group on Information Theory, 4,* 171–212.

Pike, R. (1984). Comparison of convolution and matrix distributed memory systems for associative recall and recognition. *Psychological Review, 91,* 281–294.

Qin, J., Raye, C. L., Johnson, M. K., & Mitchell, K. J. (2001). Source ROCs are (typically) curvilinear: Comment on Yonelinas (1999). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 1110–1115.

Rabin, M. D., & Cain, W. S. (1984). Odor recognition: Familiarity, identifiability, and encoding consistency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 316–325.

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

Ratcliff, R. (1988). Continuous versus discrete information processing:

Modeling accumulation of partial information. *Psychological Review, 95,* 238–255.

Ratcliff, R., & McKoon, G. (2000). Memory models. In E. Tulving & F. I. M. Craik (Eds.), *The Oxford handbook of memory* (pp. 571–581). New York: Oxford University Press.

Ratcliff, R., McKoon, G., & Tindall, M. (1994). Empirical generality of data from recognition memory receiver-operating characteristic functions and implications for the global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 763–785.

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

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

Ratcliff, R., & Starns, J. J. (in press). Modeling confidence and response time in recognition memory. *Psychological Review.*

Ratcliff, R., Thapar, A., & McKoon, G. (2004). A diffusion model analysis of the effects of aging on recognition memory. *Journal of Memory and Language, 50,* 408–424.

Reed, J. M., Hamann, S. B., Stefanacci, L. S., & Squire, L. R. (1997). When amnesic patients perform well on recognition memory tests. *Behavioral Neuroscience, 111,* 1163–1170.

Rosner, B. S., & Kochanski, G. (2008). *The law of categorical judgment (corrected) and the interpretation of psychophysical performance changes.* Manuscript under review.

Rotello, C. M., & Macmillan, N. A. (2008). Response in recognition memory. In A. S. Benjamin & B. H. Ross (Eds.), *The psychology of learning and motivation: Vol. 48. Skill and strategy in memory use* (pp. 61–94). San Diego, CA: Academic Press.

Rotello, C. M., Macmillan, N. A., & Reeder, J. A. (2004). Sum-difference theory of remember and knowing: A two-dimensional signal-detection model. *Psychological Review, 111,* 588–616.

Rotello, C. M., Macmillan, N. A., & Van Tassel, G. (2000). Recall-to-reject in recognition: Evidence from ROC curves. *Journal of Memory and Language, 43,* 67–88.

Sandusky, A. (1971). Signal recognition models compared for random and Markov presentation sequences. *Perception & Psychophysics, 10,* 339–347.

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

Schulman, A. I., & Mitchell, R. R. (1966). Operating characteristics from yes–no and forced-choice procedures. *Journal of the Acoustical Society of America, 40,* 473–477.

Sekuler, R. W. (1965). Spatial and temporal determinants of visual backward masking. *Journal of Experimental Psychology, 70,* 401–406.

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

Shipley, E. F. (1961). Dependency of successive judgments in detection tasks: Correctness of the response. *Journal of the Acoustical Society of America, 33,* 1142–1143.

Shipley, E. F. (1965). Detection and recognition: Experiments and choice models. *Journal of Mathematical Psychology, 2,* 277–311.

Simpson, A. J., & Fitter, M. J. (1973). What is the best index of detectability? *Psychological Bulletin, 80,* 481–488.

Singer, M., Gagnon, N., & Richards, E. (2002). Strategies of text retrieval: A criterion shift account. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale, 56,* 41–57.

Singer, M., & Wixted, J. T. (2006). Effect of delay on recognition decisions: Evidence for a criterion shift. *Memory & Cognition, 34,* 125–137.

Slotnick, S. D., Klein, S. A., Dodson, C. S., & Shimamura, A. P. (2000). An analysis of signal detection and threshold models of source memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 1499–1517.

Smith, D. G., & Duncan, M. J. J. (2004). Testing theories of recognition memory by predicting performance across paradigms. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 615–625.

Sorkin, R. D., & Dai, H. (1994). Signal detection analysis of the ideal group. *Organizational Behavior and Human Decision Processes, 60,* 1–13.

Staddon, J. E., King, M., & Lockhead, G. R. (1980). On sequential effects in absolute judgment experiments. *Journal of Experimental Psychology: Human Perception and Performance, 6,* 290–301.

Starns, J. J., Hicks, J. L., & Marsh, R. L. (2006). Repetition effects in associative false recognition: Theme-based criterion shifts are the exception, not the rule. *Memory, 14,* 742–761.

Stretch, V., & Wixted, J. T. (1998). Decision rules for recognition memory confidence judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 1397–1410.

Swets, J. A. (1964). *Signal detection and recognition by human observers.* New York: Wiley.

Swets, J. A. (1969). Effectiveness of information retrieval methods. *American Documentation, 20,* 72–89.

Swets, J. A. (1986a). Form of empirical ROCs in discrimination and diagnostic tasks: Implications for theory and measurement of performance. *Psychological Bulletin, 99,* 181–198.

Swets, J. A. (1986b). Indices of discrimination or diagnostic accuracy: Their ROCs and implied models. *Psychological Bulletin, 99,* 100–117.

Swets, J. A., & Birdsall, T. G. (1967). Deferred decision in human signal detection: A preliminary experiment. *Perception & Psychophysics, 2,* 15–24.

Swets, J. A., Pickett, R. M., Whitehead, S. F., Getty, D. J., Schnur, J. A., Swets, J. B., et al. (1979, August 24). Assessment of diagnostic technologies. *Science, 205,* 753–759.

Swets, J. A., Shipley, E. F., McKey, M. J., & Green, D. M. (1959). Multiple observations of signals in noise. *Journal of the Acoustical Society of America, 31,* 514–521.

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

Tanner, W. P., Jr. (1960). Theory of signal detectability as an interpretive tool for psychophysical data. *Journal of the Acoustical Society of America, 32,* 1140–1141.

Tanner, W. P., Jr., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review, 61,* 401–409.

Thomas, E. A. (1973). On a class of additive learning models: Error-correcting and probability matching. *Journal of Mathematical Psychology, 10,* 241–264.

Thomas, E. A. (1975). Criterion judgment and probability matching. *Perception & Psychophysics, 18,* 158–162.

Thomas, E. A., & Legge, D. (1970). Probability matching as a basis for detection and recognition decisions. *Psychological Review, 77,* 65–72.

Thomas, E. A., & Myers, J. L. (1972). Implications of latency data for threshold and nonthreshold models of signal detection. *Journal of Mathematical Psychology, 9,* 253–285.

Thornton, T. L. & Gilden, D. L. (in press). *What can be seen in a glance?* Manuscript under review.

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

Torgerson, W. S. (1958). *Theory and methods of scaling.* New York: Wiley.

Townsend, J. T., & Ashby, F. G. (1982). Experimental test of contemporary mathematical models of visual letter recognition. *Journal of Experimental Psychology: Human Perception and Performance, 8,* 834–864.

Townsend, J. T., & Landon, D. E. (1982). An experimental and theoretical investigation of the constant ratio rule and other models of visual letter confusion. *Journal of Mathematical Psychology, 25,* 119–162.

Trehub, S. E., Schneider, B. A., Thorpe, L. A., & Judge, P. (1991). Measures of auditory sensitivity in early infancy. *Developmental Psychology, 27,* 40–49.

Treisman, M. (1987). Effects of the setting and adjustment of decision criteria on psychophysical performance. In E. E. Roskam & R. Suck (Eds.), *Progress in mathematical psychology* (pp. 253–297). New York: Elsevier Science.

Treisman, M. (2002). Is signal detection theory fundamentally flawed? A response to Balakrishnan (1998a, 1998b, 1999). *Psychonomic Bulletin & Review, 9,* 845–857.

Treisman, M., & Faulker, A. (1984). The effect of signal probability on the slope of the receiver operating characteristic given by the rating procedure. *British Journal of Mathematical and Statistical Psychology, 37,* 199–215.

Treisman, M., & Williams, T. C. (1984). A theory of criterion setting with an application to sequential dependencies. *Psychological Review, 91,* 68–111.

Ulehla, Z. J. (1966). Optimality of perceptual decision criteria. *Journal of Experimental Psychology, 71,* 564–569.

Van Orden, G. C., Holden, J. G., & Turvey, M. T. (2003). Self-organization of cognitive performance. *Journal of Experimental Psychology: General, 132,* 331–350.

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

Verde, M. F., & Rotello, C. M. (2007). Memory strength and the decision process in recognition memory. *Memory & Cognition, 35,* 254–262.

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.

Verplanck, W. S., Cotton, J. W., & Collier, G. H. (1953). Previous training as a determinant of response dependency at the threshold. *Journal of Experimental Psychology, 46,* 10–14.

Wagenmakers, E.-J., Farrell, S., & Ratcliff, R. (2004). Estimation and interpretation of 1/f noise in human cognition. *Psychonomic Bulletin & Review, 11,* 579–615.

Wagenmakers, E.-J., Farrell, S., & Ratcliff, R. (2005). Human cognition and a pile of sand: A discussion on serial correlations and self-organized criticality. *Journal of Experimental Psychology: General, 134,* 108–116.

Wais, P., Wixted, J. T., Hopkins, R. O., & Squire, L. R. (2006). The hippocampus supports both the recollection and the familiarity components of recognition memory. *Neuron, 49,* 459–468.

Wald, A. (1950). *Statistical decision functions.* New York: Wiley.

Ward, L. M. (1973). Use of Markov-encoded sequential information in numerical signal detection. *Perception & Psychophysics, 14,* 337–342.

Ward, L. M., & Lockhead, G. R. (1971). Response system processes in absolute judgment. *Perception & Psychophysics, 9,* 73–78.

Watson, C. S., Rilling, M. E., & Bourbon, W. T. (1964). Receiver-operating characteristics determined by a mechanical analog to the rating scale. *Journal of the Acoustical Society of America, 36,* 283–288.

Wertheimer, M. (1953). An investigation of the "randomness" of threshold measurements. *Journal of Experimental Psychology, 45,* 294–303.

White, K. G., & Wixted, J. T. (1999). Psychophysics of remembering. *Journal of the Experimental Analysis of Behavior, 71,* 91–113.

Whitmore, J. K., Williams, P. I., & Ermey, H. L. (1968). Psychometric function from Rayleigh-Rayleigh ROC curves. *Journal of the Acoustical Society of America, 44,* 370–371.

Wickelgren, W. A. (1968). Unidimensional strength theory and component analysis of noise in absolute and comparative judgments. *Journal of Mathematical Psychology, 5,* 102–122.

Wickelgren, W. A., & Norman, D. A. (1966). Strength models and serial position in short-term recognition memory. *Journal of Mathematical Psychology, 3,* 316–347.

Wickens, T. D. (2002). *Elementary signal detection theory*. New York: Oxford University Press.

Wixted, J. T. (2007). Spotlighting the probative findings: Reply to Parks and Yonelinas (2007). *Psychological Review, 114,* 203–209.

Wixted, J. T., & Squire, L. R. (2004a). Recall and recognition are equally impaired in patients with selective hippocampal damage. *Cognitive, Affective, & Behavioral Neuroscience, 4,* 58–66.

Wixted, J. T., & Squire, L. R. (2004b). Recall, recognition, and the hippocampus: Reply to Yonelinas et al. (2004). *Cognitive, Affective, & Behavioral Neuroscience, 4,* 401–406.

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

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.

Yonelinas, A. P. (1997). Recognition memory ROCs for item and associative information: The contribution of recollection and familiarity. *Memory & Cognition, 25,* 747–763.

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.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46,* 441–517.

Yonelinas, A. P., Hockley, W. E., & Murdock, B. B. (1992). Tests of the list-strength effect in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 345–355.

Yonelinas, A. P., Kroll, N. E., Dobbins, I., Lazzara, M., & Knight, R. T. (1998). Recollection and familiarity deficits in amnesia: Convergence of remember–know, process dissociation, and receiver operating characteristic data. *Neuropsychology, 12,* 323–339.

Yonelinas, A. P., Kroll, N. E., Quamme, J. R., Lazzara, M. M., Sauvé, M.-J., Widaman, K. F., et al. (2002). Effects of extensive temporal lobe damage or mild hypoxia on recollection and familiarity. *Nature Neuroscience, 5,* 1236–1241.

Yonelinas, A. P., Quamme, J. R., Widaman, K. F., Kroll, N. E. A., Suavé, M. J., & Knight, R. T. (2004). Mild hypoxia disrupts recollection, not familiarity. *Cognitive, Affective, & Behavioral Neuroscience, 4,* 393–400.

Zwislocki, J., Marie, F., Feldman, A. S., & Rubin, H. (1958). On the effect of practice and motivation on the threshold of audibility. *Journal of the Acoustical Society of America, 30,* 254–262.

(*Appendixes follow*)

Appendix A

Derivation of d_e in the Noisy Decision Theory of Signal Detection

The value of d_e can be derived from the geometry of the isosensitivity space, analogously to how d_a is derived in the body of the article. We consider the point at which the isosensitivity function must intersect with a line of inverse slope through the origin:

$$z\text{HR} = -z\text{FAR}.$$

d_e is the point at which Equation 2 and this line meet, which is

$$\left(\frac{-\mu_1}{\sqrt{\sigma_1^2+\sigma_c^2+\sqrt{1+\sigma_c^2}}}, \frac{\mu_1}{\sqrt{\sigma_1^2+\sigma_c^2+\sqrt{1+\sigma_c^2}}}\right).$$

The distance from the origin to this point is

$$\text{noisy } d_e^* = \frac{\sqrt{2}\mu_1}{\sqrt{\sigma_1^2+\sigma_c^2+\sqrt{1+\sigma_c^2}}}.$$

This value is again scaled by $\sqrt{2}$ (see main text for details):

$$\text{noisy } d_e = \frac{2\mu_1}{\sqrt{\sigma_1^2+\sigma_c^2+\sqrt{1+\sigma_c^2}}}.$$

Appendix B

Demonstration That a Model in Which Criterion Variability Affects Ensemble Size Reduces to the Zero Criterial Variance Model

The two models are represented in Equations 5 and 6 in the body of the article. We start with Equation 6 and perform a little algebraic manipulation and rearrangement:

$$z\text{HR} = \frac{\sqrt{n}\mu_1}{\sqrt{\sigma_1^2+\sigma_C^2}} + z\text{FAR}\frac{\sqrt{\frac{1+\sigma_C^2}{n}}}{\sqrt{\frac{\sigma_1^2+\sigma_C^2}{n}}}.$$

$$z\text{HR} = \frac{\sqrt{n}\mu_1}{\sqrt{\sigma_1^2+\sigma_C^2}}\left(\frac{\frac{1}{\sqrt{1+\sigma_C^2}}}{\frac{1}{\sqrt{1+\sigma_C^2}}}\right) + z\text{FAR}\frac{\sqrt{1+\sigma_C^2}}{\sqrt{\sigma_1^2+\sigma_C^2}}\left(\frac{\frac{1}{\sqrt{1+\sigma_C^2}}}{\frac{1}{\sqrt{1+\sigma_C^2}}}\right).$$

$$z\text{HR} = \frac{\frac{\sqrt{n}\mu_1}{\sqrt{1+\sigma_C^2}}}{\frac{\sqrt{\sigma_1^2+\sigma_C^2}}{\sqrt{1+\sigma_C^2}}} + z\text{FAR}\frac{1}{\frac{\sqrt{\sigma_1^2+\sigma_C^2}}{\sqrt{1+\sigma_C^2}}}.$$

$$\text{Let } \sigma_* = \frac{\sqrt{\sigma_1^2+\sigma_C^2}}{\sqrt{1+\sigma_C^2}}.$$

$$\text{Let } \mu_* = \frac{\mu_1}{\sqrt{1+\sigma_C^2}}.$$

$$z\text{HR} = \frac{\mu_*}{\sqrt{\frac{\sigma_*^2}{n}}} + z\text{FAR}\frac{1}{\sigma_*}.$$

This is equivalent to Equation 5, thus showing that the two models are equivalent in form.

Appendix C

Monte Carlo Simulations and Assessment of Model Flexibility

In this appendix, we report the results of a series of simulations of the ensemble recognition task intended to assess the degree to which the noisy decision theory of signal detection (ND-TSD) spuriously captures variability that reflects failures of basic assumptions of theory of signal detection (TSD), rather than criterion variability itself. Seven simulations are reported in which data were generated assuming a failure of distribution shape assumptions (Simulations 1–4), a failure of decision rule assumptions (Simulation 5), or no such failures (Simulations 6–7). With the exception of Simulations 5 and 7 (as described in greater detail below), criterion variability was 0. Each simulation employed 30 old (signal) stimuli and 30 new (noise) stimuli per ensemble size (just as in the experiment) and 50 simulated subjects.

Failures of Distributional Assumptions

It has long been known that quite substantive departures from the assumption of normal distributions still lead to roughly linear isosensitivity functions in normal-deviate coordinates (Lockhart & Murdock, 1970). Such a finding leads to concern that the parameters yielded by a model may not accurately capture the underlying generating process and that TSD may appear to be a good explanation of the underlying decision-making process when it is not. Here, we ask whether failures of such distributional assumptions benefit ND-TSD, with the implication that the validity of ND-TSD would be undermined by providing a superior fit to data generated under alternative assumptions.

The first two models considered the possibility that the generating distributions are exponential in form. In Simulation 1, only the noise distribution was assumed to be exponential, and in Simulation 2, both distributions were assumed to be exponential. For both simulations, the rate parameter for the exponential noise distribution (λ) was set to 1. The signal distribution in Simulation 1 was normal, with mean 1 and unit variance, and the signal in Simulation 2 was exponential with $\lambda = 2$. Criteria were set to reside at a constant proportion of the average of the distributions means (.25, .50, .75, 1, and 1.25). Performance in conditions with ensembles of two and four was generated using the averaging rule.

Simulations 3 and 4 used mixture distributions for the signal distributions (DeCarlo, 2002). In both cases, the mixing parameter was 0.5. The simulations differed in the placement of the distributions; in Simulation 3, the signals were 2.5 d' units apart from one another ($d'_1 = 0.5$ and $d'_2 = 3.0$), and in Simulation 4, they were 0.5 d' units apart from one another ($d'_1 = 0.25$ and $d'_2 = 0.75$). The criteria were again set to a constant proportion of the average d' values (.28, .48, .7, .91, and 1.21). Again, ensemble performance was generated using the averaging rule.

Failures of Assumptions About the Ensemble Decision Rule

Simulation 5 investigated a case in which an alternative ensemble decision rule was used to generate the data. When criterion variance is 0, the summation rule reduces to the averaging rule, as can be seen by comparing Equations 5 and 7. Thus, for this simulation, σ_C was set to 0.8. The noise distribution was set to the standard normal distribution, and the signal distribution was set to be normal, with a mean of 1 and a standard deviation of 1.4. Criteria were set at -0.28, 0.21, 0.7, 1.19, and 1.533 and multiplied by the relevant ensemble size for the multiple-item conditions.

Standard Assumptions of TSD

In the final two simulations, the ability of TSD and ND-TSD to accurately account for data generated under their own assumptions was tested. In both, the noise distribution was the standard normal distribution. In Simulation 6, the signal distribution was normal, with a mean of 1 and standard deviation of 1.25, and there was no criterion variability. In Simulation 7, the signal distribution standard deviation was 1.4, and the criterion standard deviation was 0.8. Combining the two sources of noise in this simulation yielded approximately the same total amount of variance as in Simulation 6. In both cases, the criteria were set to -1.2, -0.5, 0.5, 1.5, and 2.2.

Candidate Models and Model Fitting

Here, we consider the performance of a number of models in fitting the data generated in the simulations described above. These are roughly the same models used to fit the actual data generated in the experiment in this article. There are eight models that represent the full combination of three factors. Each model either had a criterion variance parameter (ND-TSD) or did not (TSD). With the exception of Simulation 5, each model had either a full set of 15 criteria, five for each of the three ensemble conditions or only five criteria (restricted set of criteria). Finally, each model used either the averaging or the summation decision rule. Details of the actual model fitting are presented in Appendix D.

Simulation Results

The results of the simulations are summarized in Table C1. Across all four simulations in which distributional assumptions of TSD were violated (Simulations 1–4), TSD was much more likely than ND-TSD to achieve a superior fit. In addition, the models using an averaging rule and a restricted set of criteria outperformed their counterparts (as is appropriate, given the generating models). The lesson of these simulations is that the extra parameter provided by ND-TSD does not benefit that model in accounting for variability that derives from distributional failures of TSD.

In Simulation 5, we considered whether ND-TSD would benefit from an incorrect specification of the decision rule. Because the summation process leads to a major rescaling across ensemble sizes, it did not make sense to have equivalent criteria across ensembles. Instead, criteria were used that were a multiplicative constant across ensemble size. Thus, two additional models were fit (in the rightmost columns of Table C1 for Simulation 5) in which there were only five free parameters for criteria (like the restricted criteria models) but were multiplied by the ensemble size

Table C1
Akaike Weights for the Simulations Described in Appendix C

	Full set of criteria				Restricted set of criteria				Proportional criteria	
	Averaging rule		Summation rule		Averaging rule		Summation rule			
Simulation	ND-TSD	TSD	ND-TSD	TSD	ND-TSD	TSD	ND-TSD	TSD	ND-TSD	TSD
1	0.01	0.00	0.00	0.00	0.18	**0.80**	0.00	0.00		
2	0.00	0.00	0.00	0.00	0.18	**0.82**	0.00	0.00		
3	0.00	0.00	0.00	0.00	0.27	**0.67**	0.02	0.04		
4	0.00	0.00	0.00	0.00	0.22	**0.67**	0.04	0.08		
5	0.00	0.00	0.00	0.00	0.08	0.33	0.00	0.00	0.25	**0.34**
6	0.00	0.00	0.00	0.00	0.22	**0.78**	0.00	0.00		
7	0.00	0.00	0.00	0.00	**0.85**	0.15	0.00	0.00		

Note. Bold values indicate the winning model. These simulations reveal that ND-TSD does not benefit from excessive flexibility. ND-TSD = noisy decision theory of signal detection; TSD = theory of signal detection.

(leading to 15 different criterion values). This was the generating model, and as expected, it did outperform the other models (note that the two proportional criteria models together achieved an Akaike weight of 0.59). However, it is important to note that the simulation did not effectively recover the relatively small amount of criterion variance in the simulation (the ND-TSD version of the proportional model was outperformed by the TSD version). This result suggests that, to the degree that there is any biasing of model performance, it is toward the models without criterion variance.

The final two simulations conformed to the assumptions of TSD and ND-TSD, respectively (under the assumptions of averaging and restricted criteria). As can be seen, the model fitting successfully recovered the original model, with quite high Akaike weight scores.

Appendix D

Model-Fitting Procedures

All models were fit simultaneously to the response frequencies of individual subjects for all three ensemble sizes. Parameters were determined using maximum-likelihood estimation, as detailed below.

Criterion Variance Model and General Technique

The model predicts that the proportion of responses above the *j*th criterion, c_j, for old items in ensemble size n is

$$p(e_{\text{total}} \geq c_j|\text{old}) = \Phi\left(\frac{\mu_1 - c_j}{\sqrt{\frac{\sigma_1^2}{n} + \sigma_C^2}}\right) \quad \text{(D1a)}$$

and for new items is

$$p(e_{\text{total}} \geq c_j|\text{new}) = \Phi\left(\frac{-c_j}{\sqrt{\frac{1}{n} + \sigma_C^2}}\right), \quad \text{(D1b)}$$

where e_{total} is the total amount of evidence yielded by the ensemble, μ_1 is the mean of the signal distribution, σ_1^2 is the signal variance, σ_c^2 is the criterial variance, and n is the ensemble size. From this formula, we can derive the predicted proportion of each rating, θ_j, on the confidence scale for each item type:

$$\theta_{1j} = p(e_{\text{total}} \geq c_j|\text{old}) - p(e_{\text{total}} \geq c_{j-1}|\text{old}), \quad \text{(D2a)}$$

$$\theta_{0j} = p(e_{\text{total}} \geq c_j|\text{new}) - p(e_{\text{total}} \geq c_{j-1}|\text{new}), \quad \text{(D2b)}$$

where θ_{1j} is the proportion of the *j*th rating response for old items and θ_{0j} is the proportion of the *j*th rating response for novel items, for $j = 1, \ldots, r$, where r is the number of ratings and $c_0 = -\infty$ and $c_r = \infty$. The likelihood function for a set of parameters, μ_1, ${\sigma_1}^2$, σ_c^2, and c_j for all j, given the data, x_{ij} for all i and j, is

$$L(\mu_1, \sigma_1^2, \sigma_C^2, \text{all } c_j|\textit{all } x_{ij}) = \frac{\prod_{i=0}^{1} N_i!}{\prod_{i=0}^{1}\prod_{j=0}^{r} x_{ij}!} \prod_{i=0}^{1}\prod_{j=1}^{r} \theta_{ij}^{x_{ij}},$$

where $i = 0, 1$ indicates the new and old ensembles, respectively; N_i is the total number of the *i*th type of item; and x_{ij} is the frequency of the *j*th response to the *i*th item type. The parameter values were found that maximized the likelihood function for all three ensemble sizes jointly. Specifically, the joint likelihood function is the product of each of the three individual likelihood functions:

$$L_{\text{joint}} = \prod_{\text{all } n} L_n,$$

where L_{joint} is the joint likelihood function and L_n is the likelihood of the parameters given the data from ensemble size n. Two different sets of parameters were fit for the criterial variance model. The first had a single set of criteria, c_j where $j = 1, \ldots, r$, that was constrained to be the same for all three ensemble sizes.

The second set of parameters had $3r$ criteria, c_{jn} for $j = 1, \ldots, r$ and $n = [1, 2, 4]$, such that corresponding criteria in different ensemble sizes were free to differ. The optimal parameters were found using the mle function in MATLAB, which implements a version of the Simplex algorithm.

Zero Criterial Variance Model

The zero criterial variance model was fit by constraining σ_c^2 to be zero and maximizing the same likelihood function.

Summation Model

A version of the criterion variance model was fit in which it was assumed that information was summed, rather than averaged across an ensemble. In this model, Equations D1a and D1b are replaced with

$$p(e_{\text{total}} \geq c_j|\text{old}) = \Phi\left(\frac{n\mu_1 - c_j}{\sqrt{n\sigma_1^2 + \sigma_C^2}}\right) \quad \text{(D3a)}$$

and

$$p(e_{\text{total}} \geq c_j|\text{new}) = \Phi\left(\frac{-c_j}{\sqrt{n + \sigma_C^2}}\right). \quad \text{(D3b)}$$

All other equations remain the same.

OR Model

The OR model suggests participants perform a criterion comparison individually on each item in the ensemble and then endorse the ensemble if any of the items surpass the criterion. The model implies that the probability of endorsing an ensemble is the logical OR of the probabilities of endorsing each word. Complementarily, the probability of not endorsing an ensemble is the logical AND of not endorsing the individual words. Assuming all the words in an ensemble have the same mean and variance, on average, the logical AND of the misses (or of the correct rejections) would be probability of a miss (or correct rejection) raised to the ensemble size power. More formally, the probability of endorsing an ensemble size at a given rating level is equal to one minus the probability of none of the individual items meeting or surpassing the criterion below that rating:

$$p(e_{1 \text{ or more of } n} \geq c_j|\text{old}) = 1 - \Phi\left(\frac{c_j - \mu_1}{\sigma_1}\right)^n, \quad \text{(D4a)}$$

$$p(e_{1 \text{ or more of } n} \geq c_j|\text{new}) = 1 - \Phi\left(\frac{c_j}{\sigma_1}\right)^n. \quad \text{(D4b)}$$

Using these probabilities, the predicted proportion of each rating, θ_{ij}, for the OR model can be computed by plugging these probabilities into Equations D1a and D1b. Likelihood Equations D2a and D2b can then be used to find the maximum-likelihood estimators for this model. Like the criterial variance and zero criterial variance models, the OR model was also fit both using the same set of criteria across ensemble sizes and using a unique set of criteria for each ensemble size.

Received January 10, 2007
Revision received September 25, 2008
Accepted September 27, 2008 ■